UNIVERSITY OF OSLO
Department of Informatics

Detection of horizontal transfer events in eukaryotes using bioinformatics methods, with a focus on incorporated dietary DNA

Master Thesis

Irja Wold Sætre

May 2007

Abstract

This paper deals with horizontal gene transfer in eukaryotes, and the use of bioinformatics methods to detect such events. Horizontal gene transfer is the process where genetic material is exchanged between species. It has been known for decades that bacteria and other prokaryotes can obtain DNA from their environment and incorporate it into the genome. A potential source for horizontal gene transfer events in eukaryotes is the uptake of dietary DNA. Recent experiments have shown that uptake of DNA from feed is possible in different animals. This thesis gives an account of the horizontal gene transfer process and aims at developing a new method for detecting such events. An initial algorithm was created, but during the work with the thesis this was abandoned for the benefit of a second approach called the kTuple Frequency Based method. The kTuple Frequency Based method is a prototype approach for detecting horizontal transfer events.

Acknowledgements

First of all I would like to thank my primary supervisor Anja Bråthen Kristoffersen for her guidance and constructive feedback. She has been a great support throughout this thesis. A big thank you goes to my supervisors at the National Veterinary Institute, Arne Holst-Jensen, Knut Gunnar Berdal and Torstein Tengs, for providing good help with the biological aspect of this thesis.

I would also like to thank Kristin Wabakken Engell and Ståle Kristoffersen for their creativity and ability to help me think in new orbits. Last but not least, I would like to thank all of you who either kept my spirit up or helped me in some way during this thesis, with a special thank you to Bjørn Håkon Horpestad, Cecilie W. Lilleheil and Anette Matre for all your comments and suggestions.

Prerequisites

The readers of this master thesis should have a basic knowledge in computer science and the skill of programming, and a knowledge in biology equivalent to that obtained from an introductory course in bioinformatics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Decades ago it was proven that bacteria could acquire DNA from it's environment. The best known example of this phenomenon is perhaps their ability to develop resistance towards antibiotics. This event is known as horizontal gene transfer. All cells contain DNA, and it has always been assumed that the DNA in food and feed is completely degraded in the intestinal system. However, recent experiments have demonstrated the uptake of dietary DNA in animals (Schubbert et al., 1994, 1997, 1998; Einspanier et al., 2001; Phipps et al., 2003; Einspanier et al., 2004; Nielsen et al., 2005, 2006). As more eukaryotic genomes are being sequenced, the disclosure of any horizontal transfers from food or feed to eukaryotes becomes possible.

The title of this thesis is

> *Detection of horizontal transfer events in eukaryotes using bioinformatics methods, with a focus on incorporated dietary DNA*

The process where genetic material is exchanged between species is known as horizontal gene transfer (HGT). The term lateral gene transfer (LGT) is preferred by some, and since both refer to the same phenomenon, they are both used in this thesis. The HGT process is the opposite of vertical inheritance. The importance of HGT is that foreign genetic information is introduced into the genome, and existing components are potentially replaced with novel functionality. The field of detecting horizontal gene transfers is a relatively new one. A brief historical view of the biological background is given next. This will be continued by a review of the biology behind the

events of horizontal transfer of genetic material in chapter 2. This chapter surveys the different views on the impact of HGT events. Chapter 2 also presents different methods which can be used in discovering such events, as well as several experiments accomplished in this field.

1.1 Biology background

Humans have for centuries been wondering where we originate from, how we evolve and how everything in the world fits together. With the publication of "On the Origin of Species by Means of Natural Selection" in 1859, Charles Darwin announced a complete theory of evolution. According to Darwin, any member of a species that differs in some way that makes it better fitted to survive the competition for the means of life, will likely pass on this advantage to all it's offsprings in eternity. Today, this central concept of his theory is firmly established as the universal principle of natural selection (encyclopaedia Britannica, 2007*b*). The cornerstone of Charles Darwin's theory of evolution is the vertical inheritance, which Brown (2003) defines as the transmission of genetic material when passed on from parent to offspring. In his book Darwin has included only one drawing which represents the divergence of variants of species. This drawing is one of the first to represent the relationships of species as a tree structure, a so called phylogenetic tree. This drawing is shown in Figure 1.1.

Ernst Haeckel was a German zoologist and evolutionist who was a strong advocate for Darwinism. He proposed new notions of the evolutionary descent of man as early as in 1866. Though it was only on a theoretical basis, he suggested that the cell nucleus was concerned with inheritance. At the time, this got little attention in the scientific community (Dayrat, 2003). It was around the same time that Gregor Mendel, an Austrian botanist, began his experiments with plants which led to the discovery of the mathematical foundation of the science of genetics (Sutton, 1903).

In the late 1920s, the English physicist F. Griffith discovered that pneumococcal cells could convert from a harmless form to a disease-causing type. This "transforming principle" was proved to be heritable. In 1943, the Avery-McLeod-McCarthy experiment at the Rockefeller Institute identified that "transforming principle" as DNA (Koonin et al., 2001).

Before the 1940s, antibiotics were not used in medical practice (de la Cruz

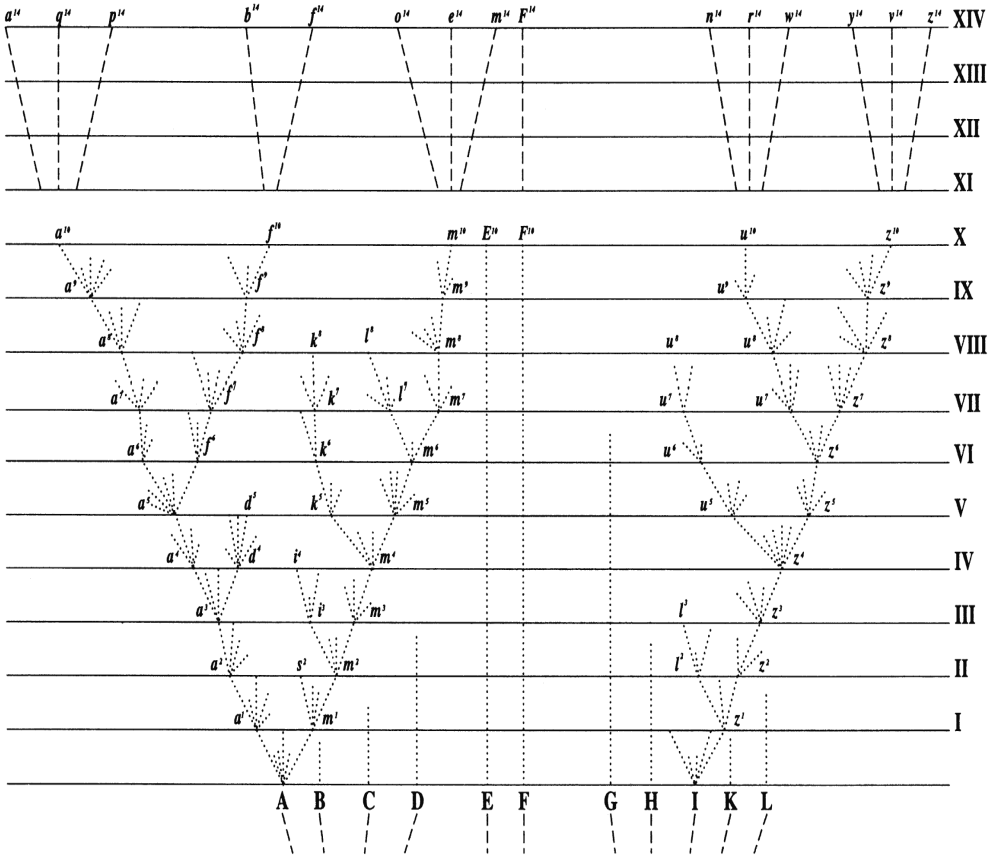


Figure 1.1: In this illustration Darwin uses the tree structure to represent the divergence of variants within a species. This is the only figure in "The Origin of Species" by Charles Darwin (L.Theobald, 2007).

and Davies, 2000). Since antibiotics became widely used, bacteria have shown the ability to develop resistance towards them. Today, this fact is one of the biggest challenges in the medical world. This phenomenom is called horizontal gene transfer, and will be described in the following chapter. Bacteria are in the family of prokaryotes, and their ability to adopt DNA from their environment has been known for decades. For eukaryotes, such as yeast, plants and animals, this phenomenon has not been experienced in the same way.

It has, until recently, generally been assumed as a fact that all DNA from food eaten when ingested is subject either to full degradation, or elimination through the feces. The first study on the fate of dietary DNA was by Schubbert et al. (1994), where mice were fed labeled DNA. The aim of this study

was to examine if dietary DNA could be taken up from the gastrointestinal tract of an animal. The authors reported that the ingested DNA survived the gastrointestinal tract, and that a proportion of 0.01% to 0.1% of the DNA fed to the mice was found to be taken up in the blood for a transient period of time. This was the beginning of continuing studies of the fate of dietary DNA in other animal species.

One of these is the study by Nielsen (2006), where tracing of dietary DNA in Atlantic salmon is one of the main goals. In this study, it was shown that dietary DNA can be taken up from the gastrointestinal tract and transported to the liver and kidney by the blood of Atlantic salmon. Other studies have documented uptake of dietary DNA in blood and organs in other animal species such as cow and calves, pigs and chicken (Einspanier et al., 2001, 2004; Phipps et al., 2003).

In the last decades genetically modified plants have been introduced for use in food and feed. With this a focus on the question of whether or not there is a potential for horizontal gene transfer of dietary DNA has risen. All food and feed contains DNA. In all organisms exposure to foreign DNA from microorganisms, plant and animal source in the gastrointestinal tract happens constantly. The dietary intake of DNA and RNA vary a great deal from organism to organism, and in humans the amount has been reported to be in the range of $0.1 - 1.0$ grams per day (Jonas et al., 2001). In average one nucleotide has a weight of 327 daltons, or atomic mass units (ambion the rna company, 2007). 1 dalton is $1.66 \cdot 10^{-24}$ grams (encyclopaedia Britannica, 2007*a*). So 1 gram of DNA corresponds to approximately $1.84 \cdot 10^{19}$ nucleotide bases. The density of nucleic acids in the cells mirrors the total amount and varies with the source. Rye and wheat from the plant family *Poaceae* contains approximately 0.7 grams of DNA per kilogram dry matter (Jonas et al., 2001). It is reasonable to assume that the DNA content of other members of this family, like maize and rice, is approximately the same. It has been reported that ordinary storage conditions have little degradation of DNA. Mechanical forces like grinding and milling have insignificantly effect on the average molecular weight of DNA, and intact DNA in crops conserved by ensiling can be detected. Food processing may lead to partial or complete degradation of DNA.

1.2 Outline of this thesis

This master thesis aims at developing a new method for detecting any DNA from feed which may have been subject for uptake and incorporation into the genome of eukaryotes during evolution. The focus in this thesis will be on comparing the sequences in the genomes of two closely related mammals in order to find short DNA sequences which only exist in one of the genomes. Then after these are found, inspect if they exist in the feed organism. Deciding the length of these short sequences will also be a part of the project. The fundamental idea and the material to use is presented in chapter 3. Chapter 4 presents the tools utilized in the implementations of this thesis. A basis algorithm on which to base the work in this thesis was initially created. This had a focus on the genomes of cow, dog and rice. The amount of data the genomes provide turned out to be a demanding challenge which caused the adoption of another approach. This second approach is called the kTuple Frequency Based method. It is a draft of a final approach, a prototype, which focuses on the genomes of cow and rice. Chapter 5 contains the different implementations accomplished in this thesis. It starts with the initial algorithm and the implementations based on this, and continues with the work that forms the kTuple Frequency Based method. This master thesis concludes with a discussion and conclusion in chapter 6.

1.3 Availability of the source code

The kTuple Frequency Based approach presented in this master thesis is a mini-version of a final approach, due to the heavy downsizing that was done. It is not implemented as a standalone program, but consists of several independent source files performing the different steps. Therefore, these different files are available on the included cd. Available on the cd is also the files holding the temporary results. The source files for the abandoned methods are not included on the cd, but may be provided by contacting the author.

Chapter 2

Horizontal Gene Transfer

In this chapter the different views on HGT is presented. A more detailed description on the HGT process as learned from prokaryotes is given. Different methods for detecting HGT events are surveyed, and several cases of HGT events are presented.

2.1 Impact of HGT

There is a general acceptance of HGT as an evolutionary factor in bacteria. Their ability to obtain genetic information from their environment is perhaps best known by antibiotic resistance. From 1941 to 1944 *staphylococcus aureus* went from susceptible of penicillin G to being able to destroy it. *Streptococcus pneumoniae*, *Streptococcus pyogenes* and staphylococci, organisms that cause respiratory and skin infections, and *Enterobacteriaceae* and *Pseudomonas* families that causes diarrhea, urinary infection and sepsis, are resistant to virtually all 160 older antibiotics (Neu, 1992). In 1941, 10 000 units of penicillin four times a day for 4 days would cure patients of pneumococcal pneumonia. Today, even with 24 million units a day a patient could die.

HGT as an evolutionary phenomenon has been the subject for an ongoing debate for a long time. Even though most scientists have acknowledged it as a major mechanism in prokaryotic genome evolution, the debate tends to become particularly energetic when lateral gene transfer cases involving eukaryotes are considered. There are two possible reasons for this reaction

(Koonin et al., 2001)

1. HGT seems to challenge the traditional vertical, tree-based view of evolution of life. It questions the core of neo-Darwinistic belief in the central role of reproductive isolation between species in evolution, concepts that have been developed in studies on evolution of sexually reproducing eukaryotes.
2. The fact that horizontal gene transfer is hard to prove unambiguously.

The latter may fade as more and more genomic sequences from eukaryotes become available, especially since recent findings indicate that HGT occurs in most lineages of eukaryotes to some degree as well as prokaryotes, and also that it is an important function in many of them (Andersson, 2005). One possible source for horizontal transfers of genetic material in eukaryotes is the uptake of DNA from food and feed.

2.1.1 Two conflicting views

It was originally thought that HGT was a phenomenon that occurred rarely. With the sequencing of more and more species, the dimension of HGT events, especially in prokaryotes, have challenged this. There is a considerable distance between the two sides in the discussion, when it comes to the extensiveness of HGT. One end saying it is such a rampant process that it is necessary to re-define the paradigm of evolution, and the other saying that it does happen in bacteria and other prokaryotes, but that it is just a cofactor. To emphasize this, when the two opposites in the HGT matter consider the same prokaryotic genome, the "enthusiastic lateralists" see genes that clearly show lateral transfer in addition to genes that show nothing clearly. The "committed verticals" on the other hand see genes that show nothing clearly, in addition to genes clearly showing vertical descent (Doolittle et al., in press).

2.1.2 The enthusiastic lateralists

The universal ancestor is the general accepted idea that all living organisms evolved from one single organism into the three main kingdoms of life, Bacteria, Archaea and Eukaryotes, that is accepted today. Woese (1998) proposes

that since the simple organisms in the very beginning experienced HGT to such an enormous extent that the universal ancestor was in fact a colony of organisms. He calls this the annealing model. This is supported in Figure 2.3, where the root is drawn as a pool.

The extreme pro-HGT wing argues that HGT is and has been such a rampant mechanism throughout history that the traditional perception of evolution needs to undergo modifications. Doolittle (1999) claims that the tree of life, where all organisms evolve from the same root, instead should be a thicket, a net, of life. An illustration of this new paradigm of life is shown in Figure 2.1.

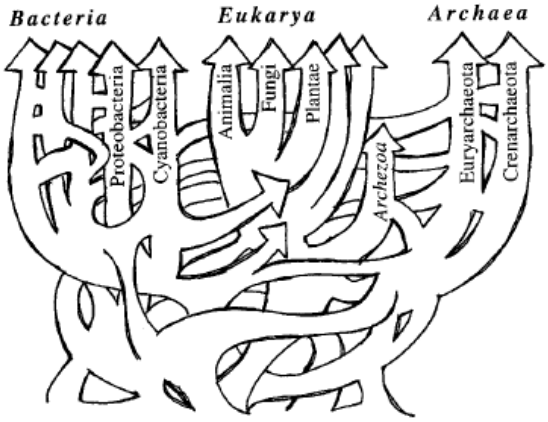


Figure 2.1: This illustration shows evolution as a thicket, or net, to account for the rampant HGT events (Doolittle, 1999).

2.1.3 The committed verticals

The other side in the debate, the committed verticals, are of the opinion that much data suggest it is an exaggeration to call horizontal gene transfer the "essence of phylogeny", and that the methods used to identify HGT events are inadequate (Kurland et al., 2003). Gogarten et al. (1999) argues that "it is often difficult to decide whether conflicts between molecular phylogenies are due to actual events in evolution, or due to artifacts generated during phylogenetic reconstruction". Although the verticals admit that HGT does occur with important evolutionary consequences, they mean it is the classical vertical inheritance that seem to be the dominant mode of evolution.

Kurland (2000) argues that in order for a transferred gene to be passed on to descendants, the transfer must be made within the germ line of the new host. For unicellar organisms this is no problem, but for higher eukaryotes, there may be barriers.

A possible and logical alternative to a horizontal gene transfer is differential loss of genes present in a common ancestor. Many independent loss events must often be invoked to produce a phylogenetic tree that could result from a single horizontal transfer. These two different explanations are illustrated in Figure 2.2. An argument against assuming differential loss as an explanation, is that each case of such an event requires the assumed presence of another gene in the genome of an ancestor. Ultimately, this means that the last universal common ancestor which lies at the presumed root of the tree of life, would have to have an enormous genome (Doolittle et al., in press).

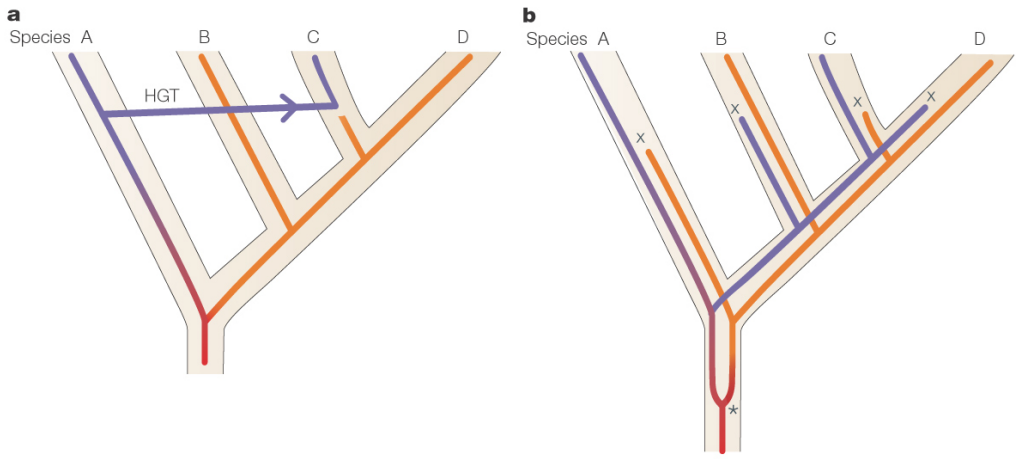


Figure 2.2: "The presence of a gene with characteristics that are typical for an unrelated group can be due to **a** Horizontal gene transfer (arrow) **b** An alternative explanation is ancient gene duplication followed by differential losses." (Gogarten and Townsend, 2005)

2.2 Steps of lateral gene transfer

There are at least four, sometimes five, distinct steps that need to occur for a gene to be transferred from one genome to another. These steps are as

presented in Smets and Barkay (2005):

1. A nucleic-acid molecule (DNA or RNA) in the donor organism is prepared for transfer.
2. The transfer step takes place. This may or may not require physical contact between the two organisms.
3. The nucleic acid enters the recipient organism through specific or non-specific means.
4. The nucleic acid is established in the recipient either as a self-replicating element or through recombination with, or transposition into, the recipient's chromosome.
5. Stable inheritance in the recipient genome might ensue.

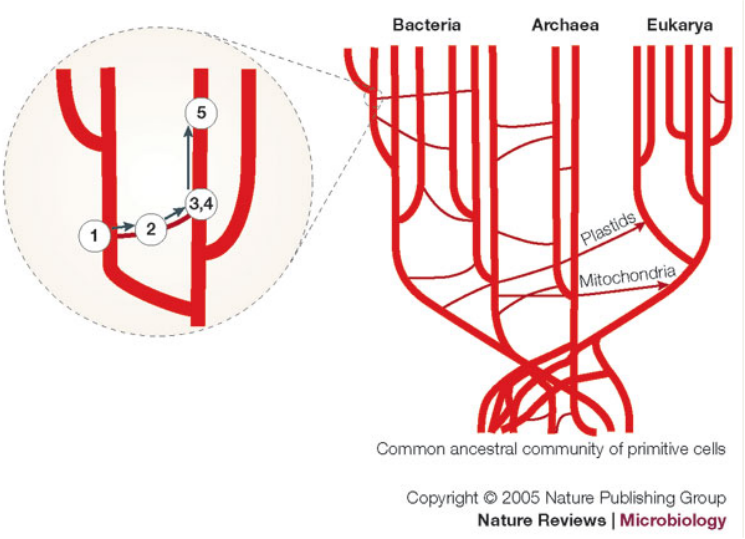


Figure 2.3: Illustration of the 5 steps of horizontal gene flow. A phylogenetic tree representing the three kingdoms of organisms. "Horizontal gene transfer and how it has impacted the evolution of life is presented through a web connecting bifurcating branches that complicate, yet do not erase, the tree of life. The inset illustrates the continuum of 5 steps that leads to the stable inheritance of a transferred gene in a new host." (Smets and Barkay, 2005)

Figure 2.3 illustrates these steps. Eisen (2000) presents the multi-step process of a HGT event in a similar manner.

A lot of reseach on horizontal gene transfer events in bacteria have discovered three mechanisms which might cause this phenomenon. These are the processes of transformation, transduction and conjugation (Licht and Wilcks, 2006).

Transduction DNA-transfer mediated by bacteriophages, which are viruses that infects a bacterium and reproduces inside it.

Conjugation Mediated either by large conjugative plasmids or by conjugative transposons, and is an active, energy-consuming process that requires cell-to-cell contact between living bacterial cells.

Transformation The process of active uptake of naked DNA from the environment and subsequent incorporation of the foreign DNA into the genome of the bacterium.

The processes of conjugation and transduction are illustrated by Figure 2.4.

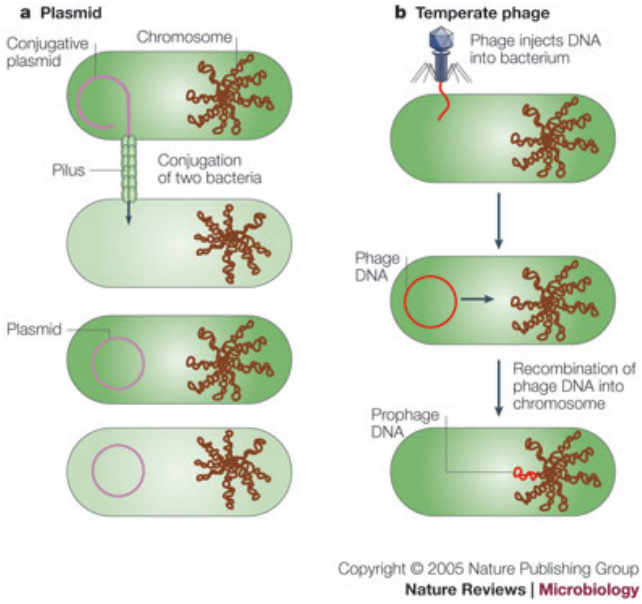


Figure 2.4: **a** Illustrates the process of conjugation. **b** The transduction process illustrated. (R.Bordenstein and Reznikoff, 2005)

Concerning horizontal gene transfer into eukaryotes, Andersson (2005) states that the critical step likely is the uptake of DNA into the cell, since these

processes probably are less wide-spread in eukaryotes. Also, it has been suggested that mammalian cells have developed several mechanisms of defense against the uptake, integration and continued expression of foreign DNA. According to Jonas et al. (2001), these defense mechanisms are thought to be:

1. degradation and/or excretion of foreign DNA
2. excision and loss of previously integrated DNA from the host genome
3. targeted inactivation of foreign genes by sequence-specific methylation

In addition, a strong barrier against the penetration of proteins and nucleic acids is the nuclear membrane. Jonas et al. (2001) states there is evidence that nuclear pores regulates the entry and that for a penetration to occur, nuclear localization signals are required. This is especially in the case of cells where the nuclear envelope is disrupted and the division have terminated.

2.3 Detection of horizontal gene transfer

There exists a variety of methods for detecting or confirming horizontal gene transfer events. These can be divided into two main groups, physical methods and bioinformatics methods.

2.3.1 Physical methods

Two complementary physical methods can identify genes that are present in some strains of a species and absent from others, as described in Doolittle et al. (in press).

Subtractive hybridization By subtractive hybridization, fragments from a target sequence that find no hybridization match in the tester DNA can be retrieved. Cloning and sequencing these fragments identifies genes that are specific to the target, which often is an unsequenced genome. This method may give false positives if the average sequence identity of the tester and the target decreases.

Comparative genomic hybridization (DNA microarrays) With comparative genomic hybridization, the process goes the opposite way. This method identifies genes that are not in the target genome. If the target and tester are too divergent to hybridize, this will give false negatives.

Examples of these techniques are Southern blot hybridization, dot blot hybridization and fluorescent *in situ* hybridization (FISH) (Feng-Guang et al., 1998; Khalsa, 2007; pasteur institute, 2007). Often used in these methods are Polymerase Chain Reaction (PCR), which is a technique used to make numerous copies of a specific segment of DNA quickly and accurately (encyclopaedia Britannica, 2007*c*). Real-time PCR is the ability to monitor the progress of the PCR as it occurs. The higher the starting copy number of the nucleic acid target, the sooner a significant increase in fluorescence is observed (applied biosystems, 2007).

2.3.2 Bioinformatics methods

There are several bioinformatics methods used to disclose horizontal transfer, as reviewed in Brown (2003); Doolittle et al. (in press); Eisen (2000).

Phylogenetic methods Phylogenetic reconstruction is considered the golden standard in HGT detection. If the evolutionary trees for many genes or proteins in many genomes are showing conflicting relationships among taxa, this might be due to HGT. Trees based on DNA sequences are best suited for studying close relatives, and protein based trees best detect ancient HGT. This is because DNA sequences change more rapidly than protein sequences. The presence of conflicting relationships among taxa can also be caused by other factors such as lineage-specific gene-losses, convergence and unequal mutation rates.

Compositional analysis Genomes differ in G+C content, codon usage and other compositional characteristics. If the nucleotide composition of neighboring sequences differ, that could indicate that a horizontal transfer has occurred. Either a gene or a non-coding region originally from another species may have been incorporated into the genome. These types of detection methods avoid some pitfalls since they only depend on the information provided by the genome, and does not rely on between-genome comparisons. But a false signal can be produced by a regional bias of the base distribution in the genome. Also, a HGT can

be hard to detect because a signal will be degraded as DNA sequences ameliorate, that is they mutate into reflecting the base composition of the new genome over time (Lawrence and Ochman, 1997).

Distribution patterns of genes If a gene is found in one species that also is present in a distantly related species, but not in close relatives, this indicates that a HGT event has occurred. Universally occurring genes will not be detected by this method. Also, gene order on the chromosome is often poorly conserved, even among close relatives. One explanation that always can be considered as an alternative to HGT for explaining this is extensive gene loss.

Homology A rapid method to detect HGT is to search for significant homologous relatives with a database search tool like BLAST. Homologous means that they share a common ancestor. A further description of this tool follows in section 4.4. When the top-scoring hits are from another taxon that is thought to be distant, a HGT event may have occurred. The similarity scores given by BLAST are not always accurately indicating the evolutionary connection. The size and type of database can affect the results. Also, it is easy to be misled by multi domain proteins.

Doolittle et al. (in press) states that both of the physical methods have been successfully used to characterize groups of closely related pathogens and "environmental microbes". As for the bioinformatic methods, it has been claimed that "each method detects preferentially certain types of horizontal transfers, and that the conjunction of several methods is required to obtain an overview of horizontal transfer extent in a genome" (Dufraigne et al., 2005). Similar conclusions were reached by Ragan (2001), who examined four different methods not requiring inference of phylogenetic trees for detecting open reading frames (ORFs) in the *E. coli* genome. These methods checked the ORFs for atypical base composition, if it could be presented by an atypical Markov model, if the BLAST matches showed an atypical pattern and if the only sure homolog in a non-adjacent lineage did not share a common phylogenetic history with the other ORFs in the genome. The four methods failed almost completely to identify a common set of *E. coli* ORFs. Ragan (2001) propose that "base composition differences may detect recent lateral transfers due to amelioration, while those focusing on cross-phylum and cross-domain patterns might detect more ancient ones."

2.4 Cases of HGT

2.4.1 Prokaryotes

Observations of HGT between prokaryotes have been observed in laboratories ever since F. Griffith in the 1920s. Since 1995 when the first complete genome of a prokaryote was available (Snel et al., 2005), several studies on detecting HGT events in prokaryotic organisms by the use of genome sequences have been accomplished. Beiko et al. (2005) performed a thorough phylogenetic analysis of $>$ 220 000 proteins from genomes of 144 prokaryotes in order to determine the contribution of gene sharing to current prokaryotic diversity, and to identify "highways" of sharing between lineages. Their results clearly show that genetic modification of organisms by lateral transfer is a widespread natural phenomenon. Dufraigne et al. (2005) studied 22 prokaryote genomes by analyzing their genomic signature, and they observed that atypical regions make up approximately 6% of each genome on average.

The sequencing of the genome of *Escherichia coli* K-12 strain MG1655 bacteria was completed in 1997 (Blattner et al., 1997). Studies by Lawrence and Ochman (1997) estimated that 17% of the protein-coding sequences of the *E. coli* chromosome were obtained through horizontal transfer. Their conclusion based on the G+C content analysis of a sequenced region including more than 30% of the *E. coli* chromosome. These findings are supported by another study by Lawrence and Ochman (1998), where it was found that 17.6% of ORFs have been introduced into the *E. coli* genome in at least 234 lateral transfer events since this species diverged from the Salmonella lineage 100 million years ago. This was found by identifying atypical GC content at first and third positions of protein coding regions.

The major enzyme of all photosynthetic cells is ribulose-l,5-bisphosphate carboxylase/oxygenase, known only as rubisco. Delwiche and Palmer (1996) found 6 horizontal gene transfers involving rubisco genes implied by their phylogeny. 2 between cyanobacteria and proteobacteria, 2 between proteobacteria and plastids, and 3 within proteobacteria. Another example of a prokaryote that is found to have experienced HGT is *Thermotoga maritima*, a bacterium which has been placed as one of the deepest and most slowly evolving lineages in the Eubacteria by small-subunit ribosomal RNA phylogeny (TIGR, 2007). Nelson et al. (1999) states that 24% of the *T. maritima* genes are most similar to archaeal genes, not those of other thermophilic bacteria. They got 81 archaeal-like genes grouped together in 15

regions of the *T. maritima* genome that range in size from 4 to 20 kilobases, when using the coding-analysis program GLIMMER. GLIMMER uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from non-coding DNA (glimmer mobile gene-finding system, 2007).

Garcia-Vallé et al. (2000) developed a statistical procedure for predicting if genes in a complete genome have been acquired by HGT. It was based on the analysis of G+C contents, codon usage, amino acid usage and gene position. When applied to 17 complete bacterial genomes and 7 archaeal ones, the percentage of horizontally transferred genes varied from 1.5% to 14.5%. They found that informational genes were less likely to be transferred than operational ones. Other studies reached similar conclusions as to what types of genes that were more likely to be transferred. Nakamura et al. (2004) developed a method based on Bayesian inference with training models for nucleotide composition for detecting horizontally transferred genes and their possible donors. They found that 14% of open reading frames in 116 prokaryotic complete genomes were subject to recent horizontal transfer. Based on their data set they quantitatively determined that the biological functions of horizontally transferred genes are biased to three categories, and that the transferability of genes seems to depend heavily on their functions. Another study by Shi et al. (2005) also found a biased distribution of the different functions of the horizontally transferred genes they observed in 82 genomes.

2.4.2 Eukaryotes

According to de la Cruz and Davies (2000), eukaryotes possess the same capacity and similar mechanisms for effective HGT as prokaryotes do, and also that laboratory experiments have shown that these mechanisms are functional. Plant mitochondria possess an active DNA uptake system, and in all but one of the 40 cases of plant-to-plant HGT reported thus far, the transferred gene is a mitochondrial gene (Richardson and Palmer, 2007). The usage of gene modified organisms in the food industry has risen the question if it is possible for dietary DNA to be taken up into the organism. Several studies have shown the uptake of dietary DNA in vertebrates.

Fish

Nielsen et al. (2005) investigated the uptake of dietary DNA into blood, kidney and liver of Atlantic salmon. The feed that was force-fed to the fish contained 3 PCR amplified DNA fragments with a high copy number. At intervals from 1 hour to 64 hours after the force feeding, tissue samples were collected. The presence or absence of the DNA targets were determined by the use of the real-time PCR method. Uptake of dietary DNA was observed. The longest fragment of 624 bp was found in the distal intestine, which is the last part of their digestive system. In the kidney and liver samples fragments of 282 bp were the longest detected. The highest concentrations of dietary DNA in liver and kidney was found 8 hours after the force feeding. In a follow up study, Nielsen et al. (2006) studied the transport in blood and accumulation in some organs of intravenously injected DNA in salmon. This DNA was detected in blood, liver, kidney and gonad up to 24 hours after injection. Targets up to 283 bp were detectable in blood and liver samples 1 hour after injection, and at the same time fragments up to 436 bp were detectable in the kidney and muscle samples. Up until 40 minutes after the iv injection, fragments up to 624 bp were detectable in the blood samples. A relatively rapid elimination of the DNA was indicated by the occurrences of longer targets at 5 minutes after injection than after 2 hours. The amount of DNA injected into the Atlantic salmon was estimated to be no more than 0.05% of what would normally be absorbed from the gastro intestinal tract during one day. Based on data from both experiments, Nielsen et al. (2006) estimate that approximately 1% of dietary DNA is absorbed in Atlantic salmon.

Mice

Schubbert et al. (1994) approached the question "Is the epithelial lining of the mammalian gastrointestinal (GI) tract a tight barrier against the uptake of ingested foreign DNA or can such foreign DNA penetrate into the organism?". In their experiment, mice were either pipette-fed bacteriophage M13 DNA or the food was added M13 DNA. The fecal of these mice had previously shown to be free of any M13 DNA, and this test molecule contains no DNA homologous to the mouse genome. The feces of the animals was tested for M13 DNA sequences at several different times after feeding, by Southern or dot blot hybridization, or by the ploymerase chain reaction (PCR). M13 DNA fragments with sizes in the range of < 200 and 400 bp were the majority

on the Southern blot hybridization. DNA from blood were also extracted at different times after the feeding, and examined for the presence of M13 DNA by dot blot hybridization or PCR. Fragments of up to 472 bp were found by PCR. M13 DNA was only detected in the feces or in the blood from 1 hour after to 7 hours after feeding the mice. The experiments showed that food-ingested foreign DNA is not completely degraded in the GI tract of mice, and suggested that 0.01 to 0.1% of the M13 DNA fed could be retrieved from the blood.

An experiment to follow up these results was performed (Schubbert et al., 1997). M13 DNA fragments were found in blood and in the nuclei of peripheral leukocytes from 2 to 8 hours after feeding, and in the nuclei of spleen or liver cells up to 24 hours after feeding. As an extension to both these experiments, Schubbert et al. (1998) studied the offspring of mice that during pregnancy were fed a daily dose of 50 micro gram of M13 DNA. DNA fragments from M13 were discovered in various organs of fetuses and newborn animals by using the PCR or fluorescent *in situ* hybridization (FISH) method. The M13 DNA fragments had a length of maximally 830 bp. In the conclusion, the authors propose that since some of the DNA is associated with chromosomes, it is probably integrated into mice DNA. They express that germ line transmission probably would have been difficult to detect under the experimental regimen chosen, and that distribution of foreign DNA in the offspring organisms are consistent with transplacental transfer. A transplacental transfer is a transfer through the membrane that surrounds the fetus in mammals.

Farm animals

Einspanier et al. (2001) studied the fate of ingested recombinant plant DNA in cattle and chicken. These farm animals were fed a diet containing conventional maize or gene modified (GM) *Bacillus thuringiensis* toxin-maize (Bt-maize). Physical methods were used. Data indicated that DNA fragments, < 200 bp, from plant choloroplasts could be detected in the blood lymphocytes of cows. Also, faint signals of plant DNAs were found in their milk. No signals were detected in blood and feces of cattle. As for the chickens, samples from all tissues examined, which were muscle, liver, spleen and kidney, DNA fragments from the maize chloroplast were detected. A study on the fate of DNA material from feed during the passage through the different parts of the cow GI system was performed in Einspanier et al. (2004).

22 cattle were fed Bt maize and different plant genes for 4 weeks, and the Bt toxin gene were analyzed. Quantitative real-time PCR was the main analytical method used for detection. This study was also the first investigation for any GM-feed induced events of the rumen bacterial population. Remarkable concentrations of the Bt toxin gene were found in all GI tract samples, and there were also traces in feces. The experiments were not able to reliably quantify low-abundant genes in the GI tract of cows using real-time PCR.

The study accomplished by Phipps et al. (2003) aimed at determining the presence or absence of tDNA and plant DNA in lactating dairy cows. The examination of ruminal fluid, duodenal digesta, feces, blood, and milk were accomplished. Genetically modified maize was fed the cows, as in the other studies, in addition to (GM) soybean meal. As in the above studies, the PCR method was used. The chloroplast rubisco gene has a great number of copies per plant cell. In the majority of the analyzed samples, fragments of the rubisco gene was detected in both ruminal and duodenal digesta, milk and feces, but rarely in blood. The findings of DNA fragments in fecal samples is in contrast to the results in (Einspanier et al., 2001). The size of the rubisco gene fragments detected in ruminal and duodenal digesta was 1176 bp long, but was decreased to 351 bp in the fecal samples.

Humans

Salzberg et al. (2001) analyzed the human genome for evidence that genes had been horizontally transferred into the genome from prokaryotic organisms. Protein sequence comparisons of human, fruit fly, nematode worm, yeast, mustard weed, eukaryotic parasites and all complete prokaryotic genomes were performed. In this study, about 40 genes were found to be exclusively shared by humans and bacteria. The "you are what you eat"-hypothesis by Doolittle (1998) proposes a genetic ratchet as a mechanism for explaining the occurrence of such microbial genes in eukaryotic organisms.

For DNA which has been incorporated into the genome to be passed on to a descendant it has to be present in the genetalian cells. Nielsen et al. (2006) provided evidence for the transportation of iv injected DNA to the gonads in Atlantic salmon, and Waters (2001) presents evidence for a conjugation from *E. coli* to Chinese hamster ovary cells, performed in a laboratory experiment.

Most studies on eukaryotic HGT has been accomplished by the use of physical methods. As Richardson and Palmer (2007) points out, PCR-based studies

have revealed much about HGT, and the next major step is to sequence whole genomes of eukaryotes known to have experienced HGT. Genome sequencing will uncover transfers that are too short or from such evolutionary distant donors that they fail to amplify by PCR with certain primers. With the availability of more genome sequences, the use of bioinformatics methods to detect HGT events in eukaryotic genomes will probably increase drastically.

Chapter 3

Fundamental idea and material to use

3.1 The fundamental idea

The previously presented examples of studies show that the uptake of DNA from feed is possible in vertebrates, and they presented possible HGT events in eukaryotes. The mechanisms for the incorporation of this DNA is present in all cells, and transfer of DNA between cells happens. An interesting task is to disclose if such events have occurred in certain organisms, and can be detected by examining genome sequences. By comparing the DNA sequences

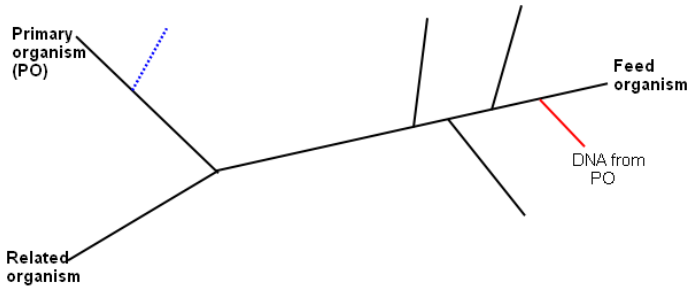


Figure 3.1: Illustration of the starting point.

from two closely related organisms, the subsequences that are specific to one of them, and therefore have evolved after the two organisms split, can be determined. After these species-specific subsequences are identified, a

search in a database can be performed to see if they are found in the genome sequences of other species. If the result of a search gives matches in a species whose lineage is thought to be divergent, this may imply that a horizontal transfer event has occurred. This idea is illustrated in Figure 3.1. In this figure, the blue line shows where a DNA sequence from the primary organism presumably would fit in an unrooted phylogenetic tree. A sequence from the primary organism which fits at the red line could indicate that a horizontal transfer has occurred.

3.2 Material

The species to use as the primary object should be one which has had a unilateral diet through all times that also preferably is fully sequenced. Fully sequenced higher order eukaryotes, such as mammals, are few and two of those, human and mouse, both have had an extremely varied diet from the beginning. Even though there are not that many fully sequenced animals several sequencing projects have been going on for several years. Hence there are sufficient material available for use.

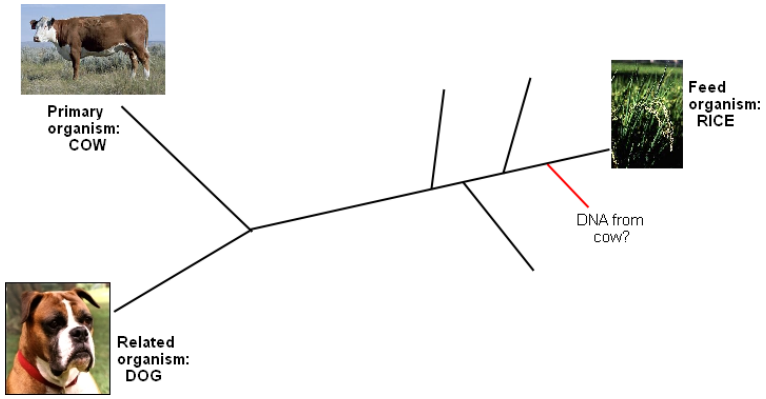


Figure 3.2: Illustration of the animals to use as material in this thesis. The photos are taken from the NCBI (2006*a*) homepage.

The animal chosen as primary is cow. The cow, *bos taurus*, genome project is on-going. Cows have from the beginning of time had a diet consisting of vegetables, mainly grass-based. A "close" relative to cow is dog. The dog, *canis familiaris*, genome project is also on-going. Dogs are meat eaters, so if a HGT event from feed to cow has occurred after cow and dog split, this should be detectable as a sequence motif only present in cow and some plant,

and not in dog. Rice, *oryza sativa*, is a plant that is fully sequenced. It is a member of the grass family called *poaceae*, which includes maize, wheat, barley and rye, among others. Rice is chosen to be used as the feed DNA source. This is illustrated in Figure 3.2.

The reference sequences of the genomes for cow and dog, and the complete rice genome were downloaded from the FTP site at the NCBI home page (NCBI, 2006*b*) in September 2006. Table 3.1 holds information on the downloaded genome sequences.

	bos taurus	*canis familiaris*	*oryza sativa*
Project finished?	No	No	Yes
Files dated	2005/10/4	2005/09/01	2006/04/15
Genome size	3000 Mb	2400 Mb	390 Mb
Chromosomes	29	38	12

Table 3.1: Information on the genomes. *Bos taurus* and *canis familiaris* are in addition to the given number of chromosomes composed by 2 sex chromosomes and the mitochondria chromosome. *Oryza sativa* is composed of chloroplast and mitochondria chromosomes in addition to the given number. The genome sizes are approximated numbers, where M=million and b=bases.

Chapter 4

Informatics background

Several programs exist that offers methods which can be used to detect HGT events in or between sequences. Some uses the phylogenetic approach, some uses different implementations of nucleotide base composition analysis, and others uses a combination of these and other methods. Existing programs will not be reviewed here.

For this thesis the Python programming language is used (Python, 2006). Some explanatory words on terms used in the implementations are necessary. A *module* is an implementation which may be run as a standalone program, or which offers methods to other implementations. A *dictionary* is what is called a hash map in other scripting and programming languages. One piece of information regarding the handling of files, is that after several tests the author has concluded it impossible to both read and write to one file in the same operation.

The implementations accomplished in this project have utilized some existing tools that are freely available. A description of these follows next.

4.1 eGenix mxTextTools

eGenix mxTextTools are fast text manipulation tools for Python that use a suffix skip array, called Fast Search Algorithm (eGenix.com Software, 2006). It is an enhanced version of the well known Boyer-Moore search algorithm,

which finds a pattern without looking at all of the characters it scans past. The longer the pattern is, the faster it moves on average (Moore, 2007).

Two of the search tools offered in this package are used in the implementation. These are

find(text, [start=0,len_text=len(text)]) Search for the substring match in text, looking only at the slice [start:len_text] and return the index where the substring was found, or -1 if it was not found.

findall(text,what,start=0,stop=len(text)) Returns a list of slices representing all non-overlapping occurrences of what in text[start:stop]. The slices are given as 2-tuples (left,right) meaning that what can be found at text[left:right].

4.2 The FASTA format

All the DNA-sequences for a chromosome are in separate FASTA files. The FASTA format is a widely used text-based format to represent nucleotide or protein sequences. It consists of a header line which is a single-line description, followed by the sequence representation. One such FASTA file can contain several sequences, as illustrated in Figure 4.1.

```
> seq1|The first sequence
TTCCCCAAGAACTTCACCATACGTCTTTCTCGTTTTGGTGCCCAGCTGAGGCAAGAGTGA
CGACTGGCTGGTGTGGTCCTGCAAGAAAGATGATGCGCCTGT
> seq2|The second sequence
GAAAGAAAAAGCATCATCTTGGCTCTTGCTTCAAAAATTATGTTATACCAAACTCCATGC
TTTCTTAATATTGTTTCTGTTCTGTGCTATTCTTTGTCGCTCAGTCATGTCCAACTTTTT
...
```

Figure 4.1: An example of a FASTA file.

4.3 Biopython

The Biopython Project is a freely available python project that offers tools for computational molecular biology (biopython Project, 2006*a*). Behind it is an international association of developers. The Biopython package includes:

1. Interfaces to common bioinformatics programs such as Standalone Blast from NCBI and Clustalw alignment program.
2. Code for dealing with alignments, including a standard way to create and deal with substitution matrices.
3. A standard sequence class that deals with sequences, ids on sequences, and sequence features.
4. The ability to parse bioinformatics files into python utilizable data structures, including support for the most commonly used formats.
5. Several other useful methods.

For this thesis especially two of the modules from the Biopython package are of great use. These are (biopython Project, 2006*b*)

Bio.Fasta Utilities for working with sequences formatted in the FASTA format. Among others, it provides an Iterator which returns one record at a time from a FASTA file. The information this record holds is the title line and the sequence itself.

Bio.pairwise2 Implements pairwise sequence alignment using a dynamic programming algorithm. It provides functions for both local and global alignments. When performing an alignment, it is possible to specify the match/mismatch scores directly as scores or by specifying a dictionary containing the scores. It is also possible to specify what kind of gap penalties should be used or to use no penalty for gap.

4.4 BLAST

BLAST (Altschul et al., 1997) is one of the major heuristic algorithms for performing database searches. It stands for Basic Local Alignment Search

Tool. The main idea is to find high scoring ungapped segments among related sequences. A threshold is set, and any segment with values above this threshold indicates pairwise similarity beyond random chance. This helps to select related sequences from unrelated sequences in a database. Recently an improvement of the BLAST implementation is the ability to provide gapped alignments (Xiong, 2006).

A family of programs implemented on the basis of the BLAST algorithm exists. These are specialized versions of BLAST:

BLASTP Searching with protein sequences in a database of protein sequnces.

BLASTN Nucleotide sequences as queries in a database of nucleotide sequences.

BLASTX Nucleotide sequnces as queries and translate them in all six reading frames to produce translated protein sequences, which are searched for in a protein database.

TBLASTN Uses protein sequences, translates them in all six reading frames for searching in a nucleotide sequence database.

TBLASTX Nucleotide sequnces that are translated in all six reading frames for searching in a database of nucleotides that also are translated in all six reading frames.

All of these programs are available from the BLAST web server on the NCBI homepage (NCBI, 2007).

There are several databases which can be used in a BLAST search. One of these are the nr database. The name of this database stands for non-redundant, which was the original idea behind it. This is no longer true, though it is the largest nucleotide database available through NCBI BLAST.

A run of a BLAST search generates an output with the results from the search. Such results will be given in section 5.8. A list of all the matches for the query sequence is given, with corresponding values. The most interesting value is the *E-value*.

4.4.1 E-value

BLAST compares a query sequence against all sequences in the database. The E-value is a statistical indicator for the likelihood that a given sequence match is given by chance. An empirical interpretation of the E-value provided by BLAST as given in Xiong (2006), is shown in Table 4.1.

E-value	**Interpretation**
$\texttt{E} < 1 \cdot 10^{-50}$	Extremely high confidence that the database match is homologous
$1 \cdot 10^{-50} < \texttt{E} < 0.01$	The match can be considered as a result of homology
$0.01 < \texttt{E} < 10$	The match is not significant, but may hint a remote homologue
$\texttt{E} > 10$	The sequences are most likely unrelated

Table 4.1: Empirical interpretation of the E-value.

An evolutionary relationship between two sequences remains constant. Because the E-value is proportionally affected by the size of the database, as the database grows the E-value for a given match of two sequences increases. The *query coverage* value is another worth taking into consideration. This value reveals how much of the query sequence is matched. The value in *Max. identity* gives the percentage of exact hits in the match. The BLAST output uses another statistical indicator in addition to the E-value, the *bit score*. This score measures the sequence similarity independent of the length of a query sequence and the size of the database. It is normally based on the raw pairwise alignment score. The higher the bit score, the more significant the match is. In this thesis only the E-value is considered when conclusions about homology are made.

Chapter 5

Implementation

A typical gene is in the range of 300 bases to 1500 bases, though both longer and shorter genes exist (Xiong, 2006). As previously described, studies on the uptake of dietary DNA provided evidence for the uptake of sequences having a length in that range into blood and organs of different animals. Most of the previously cited literature and the detection methods they used had focus on the transfer of whole genes. In this thesis, the possibility that only fractions of the DNA in a gene has been subject to a transfer is taken into consideration.

The approach adopted in this thesis is inspired by the nucleotide compositional methods. The initial approach is as follows: Collect all available DNA sequence material of an animal which through evolution has had an unilateral diet. Compare the DNA of a close relative to that of the primary organism regarding segments of a certain length. These segments are called ktuples. Remove every ktuple that is found in both the primary and the relative. The remaining ktuples will be compared to the DNA of the feed in order to find similar ktuples. The ktuples that are found in both the feed DNA and the primary DNA are the interesting ones. These should be further investigated as they might confirm the existence of HGT.

The basis idea for the initial implementation is this algorithm:

1. Make a list (hash map) of all the ktuples in cow, including the reverse complement, with k=30 as a start value. Different values for k should perhaps be tested.

2. Remove from the list all the ktuples also found in dog.
3. Check if any of the remaining ktuples exist in the plant (rice). If there are any, study these hits closer with alignments to dog/cow/plant, etc.

The usage of ktuples of length 30 entail that an enormous number of possible different ktuples have to be managed. The exact number is

$$4^k = 4^{30} = 1152921504606846976 = 1.15 \cdot 10^{18}$$

The total number of bases in the cow files is $3100376884 = 3.1 \cdot 10^9$. The probability for all the different ktuples to be present is

$$\frac{3.1 \cdot 10^9}{1.15 \cdot 10^{18}} = 2.7 \cdot 10^{-7}\%$$

which is a very low probability.

5.1 HGT_nr1

Some of the properties an implementation needs are handle FASTA files, accumulate a large number of ktuples of a certain length and produce the reverse complement of a DNA sequence. The basis idea was implemented and called HGT_nr1. In this first implementation, the algorithm is as follows:

1. Fetch all different overlapping ktuples present in cow, including the reverse complements.
2. Fetch all overlapping ktuples present in dog.
3. Delete all ktuples in cow that are also present in dog.
4. Fetch all overlapping ktuples present in plant.
5. Check if the remaining cow ktuples occur in plant.
6. Treat the remaining ktuples further.

All overlapping ktuples found are stored in the internal memory in a separate dictionary for each organism. The understanding of overlapping ktuples is

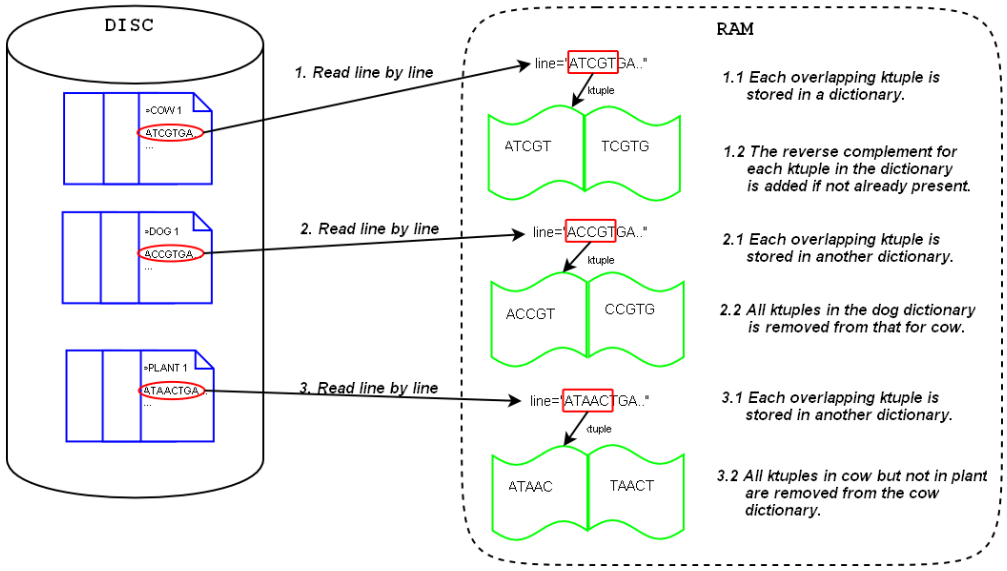


Figure 5.1: An illustration of the HGT_nr1-method. The green "books" with one ktuple on each page in the dotty RAM area are illustrating the dictionaries holding each ktuple as the key-value. Overlapping ktuples means that one ktuple starts in position x and the next starts in position $x + 1$.

that if one ktuple starts in position x, the next ktuple starts in position $x+1$. The contrary is non-overlapping ktuples, where one ktuple starts in position x and the next starts in position $x + k$. An illustration of this approach is shown in Figure 5.1

A test-run of this implementation was executed on a computer with 1GB RAM and an Intel®Pentium®4 3.2GHz processor, having the Microsoft Windows XP©operating system. The files used in this run had sizes equal to just a fraction of the full genome files, and it took several minutes to complete. There was no problem handling files with small sizes like this, but as the files got bigger, the trouble became obvious. When a cow file of approximately 116 MB was used, the program caused the machine to crash. The length of the sequences in this file is approximately $122 \cdot 10^6$ bases. The same test-run was performed on a computer with 6 GB RAM and 4 Intel®Xeon^{TM}CPU 3.60 GHz having the Linux operating system. This run was abnormally completed with a "MemoryError"-message.

5.2 HGT_nr2

A change in the implementation was necessary in order to get an algorithm that functions with big files. In the previous attempt too much data was kept in the internal memory. This lead to a memory error. A new and different approach was implemented. In this method the goal was to not exhaust the RAM. The basic idea of finding the ktuples in cow which are not in dog was applied. After these ktuples were found, a search for them in rice was executed. A ktuple length of 30 was still used, but a more efficient method for creating the reverse complement sequence was implemented. The latter used a built-in method to get the reverse of a string. This operation is not affected by the size of a sequence, and thereby performs in a constant, near zero, time. The new algorithm wass called HGT_nr2. The steps in it can be summarized as:

1. Scan through the cow-genome for ktuples.
2. When a new ktuple is found, check if it is present in any dog sequence.
   (a) if yes, go to next ktuple, start from 2.
   (b) if no, check if present in any plant sequence.
      i. if yes, save in file.
      ii. if no, continue with next ktuple.
      iii. do the same for reverse complementary ktuple.

This method is illustrated in Figure 5.2.

A test-run of this implementation showed that it was extremely slow, even for small files. A profiling of an execution revealed that 99% of the time was used in the exist-method. The exist-method is where a file is opened, read to memory and a search for the ktuple is done before the file is closed. As is obvious from the pseudo code presented in Figure 5.3, an unacceptable number of I/O-operations take place even with small files. I/O-operations are time consuming compared to operations performed in the internal memory.

The HGT_nr2 algorithm is in a way the opposite of the HGT_nr1, where too much data is kept in memory at the same time. In this one, little data is kept in memory while too many I/O-operations are necessary.

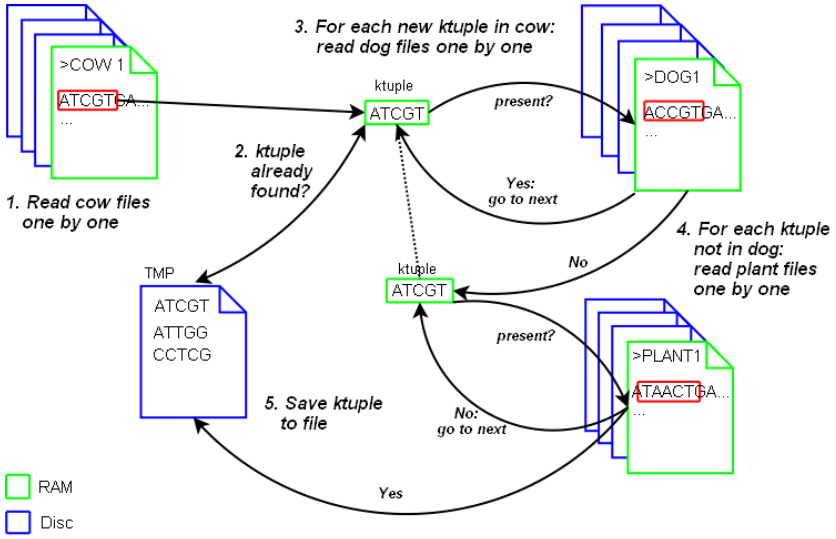


Figure 5.2: An illustration of the HGT_nr2-method. The figures drawn in a blue color symbolizes being stored on the disc, while figures drawn in a green color are kept in memory.

```
 1 for every_pos_in_seq:
 2   ktup=seq[i:(i+k)]
 3   if exist(ktup,ktup_file):
 4       continue
 5   if exist(ktup,dog_file):
 6       continue
 7   if exist(ktup, plant_file):
 8       save ktup in ktup file
 9   revtup=make_rev(ktup)
10   if exist(revtup,ktup_file):
11       continue
12   if exist(revtup,dog_file):
13       continue
14   if exist(revtup, plant_file):
15       save revtup in ktup file
```

Figure 5.3: Pseudocode for the method which finds the ktuples in the HGT_nr2 implementation. The amount of open-read-check-close-operations is obvious.

5.2.1 HGT_nr2 version 2

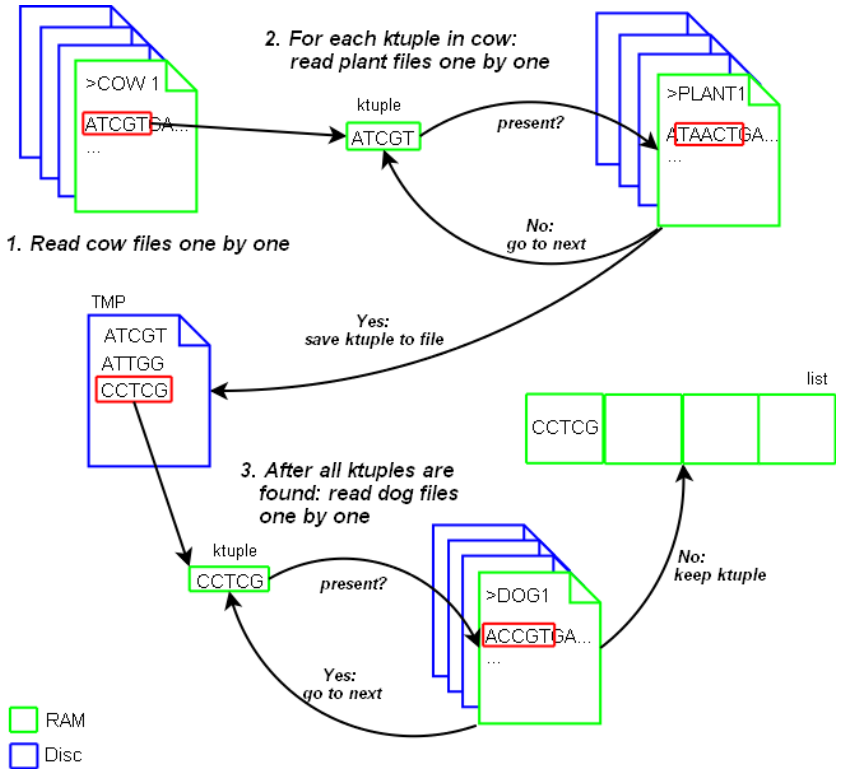


Figure 5.4: An illustration of the HGT_nr2-method version 2.

The genome of *canis familiaris* and *bos taurus* should theoretically have more in common than *bos taurus* and *oryza sativa*. In order to cut down the amount of I/O-operations, it was decided to change the order in which the tasks are accomplished. The new succession was first to compare cow and rice, and then use dog to filter the outcome. Another improvement compared to the original HGT_nr2 method was the use of the mxTexttools package for checking if a file contains a ktuple. This revised version of the HGT_nr2-method is illustrated in Figure 5.4. The refinements lead only to minimal improvements. The algorithm was still too time consuming.

5.3 HGT_nr3

The previous attempts to scan through the genomes for 30 bp long overlapping segments, and either store them in memory or in a file turned out

not to work. It is not possible to keep all the different ktuples of length 30 present in cow in memory at once, and keeping them in a file was too slow the way it was implemented. If shorter sequences were to be considered, it would be possible to keep them all in memory while working with them. This would lead to less significant biological results, since statistically it is more likely that at short sequence is found than a longer one. When comparing an approach which fails to execute with one that provides an approximated answer, the latter wins.

An approach quite different from the previous ones was proposed. The new idea was to divide the ktuples of length 30 into shorter ones, keep a survey over their positions and later connect them into the wanted length. A focus on those occurring in the plant genome was proposed, and later perform searches for these in the cow genome. This approach would still result in handling ktuples of length 30 in the end. Though this approach was abandoned before it was implemented, it inspired a new one.

Moving the focus back to the cow genome as before, another way to seek for ktuples which possibly origin from plant was formed. The theoretical assumption was made that if a horizontal transfer from rice to cow is thought to have occurred, it is reasonable to believe that the number of times the transferred DNA segment is present in the cow genome would be low compared to original cow DNA segments. This is stated as the hypothesis

> *A DNA sequence which is horizontally transferred from feed to animal and is incorporated into the animal's genome, occurs rarely.*

This view on horizontally transferred DNA is supported in Phipps et al. (2003) which states that "if plantDNA is absorbed [...] the frequency is likely to be exceedingly low."

The new approach was to use the frequencies of ktuples with a length < 30 in the cow genome to find those that may have been transferred horizontally. Later the ktuples should be expanded to get more significant biological results when aligning them. BLAST searches should be executed to verify the origin of the segments in the alignments.

5.4 The ktuple distribution in cow

The first step was to decide the length of the ktuples to work with. The results from the previous approaches entail a length < 30 bp. The total number of bases in the cow-files is $3.1 \cdot 10^9$. If an even distribution is assumed, each ktuple of length k is observed

$$\frac{3.1 \cdot 10^9}{4^k}$$

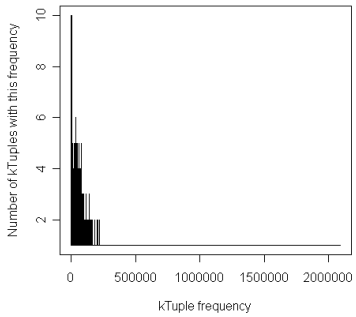
times in the cow genome. It is most unlikely that the ktuples are evenly distributed. To get an idea of the distribution of ktuples of different lengths in cow, all ktuples of lengths 8, 10 and 12 were gathered, along with their frequencies. Information on these ktuples is shown in Table 5.1. A module performing this task was implemented. It was obvious that the distribution

k	4^k possible ktuples	All found?	$\frac{3.1 \cdot 10^9}{4^k}$	Different frequencies found
8	65 536	yes	47302	43871
10	1 048 576	yes	2956	25066
12	16 777 216	no	184	16893

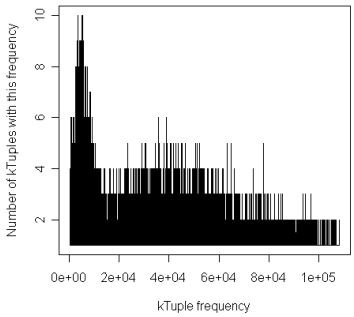
Table 5.1: Information on the ktuples in cow.

of the ktuples was even for neither of the lengths. Looking at one ktuple and how many times it occurs provides limited infomation when a survey of the distribution is wanted. More interesting is to see how the number of times a ktuple occurs is compared to the frequencies of the other ktuples. If for instance a ktuple is observed 100 times, is this a small or big number compared to the other ktuples. To get an overview of this all the observed frequencies and the number of ktuples having that frequency was plotted against each other in "kTuple frequency vs. Number of ktuples with this frequency"-plots for each value of k.

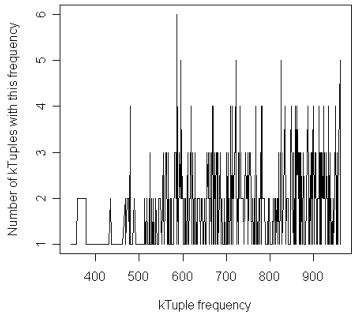
Figure 5.5 shows the distribution of ktuples in the cow genome for k=8 in three different levels. All possible ktuples are observed at least once, and the frequencies vary from over 2 million to approximately 300. The number of ktuples having these frequencies are in the range from 1 to 10. The distribution of ktuples in the cow genome for k=10 is shown in Figure 5.7. As for k=8, all ktuples of length 10 are observed. The frequencies ranges from 3 to approximately 1 million, whereas the number of ktuples having these frequencies are approximately 1000 and descending. Figure 5.6 shows the distribution of ktuples in the cow genome for k=12. Around



(a)All the observed frequencies.



(b)A zoom in on the area having the most frequencies.



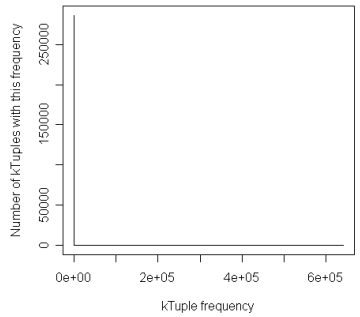
(c)Further zoom in on the lowest observed frequencies.

Figure 5.5: Plots showing the distribution of ktuples in the cow genome for k=8 in three levels.

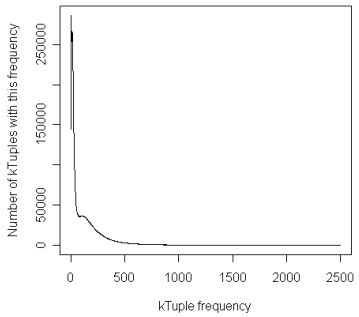
150 000 of these ktuples are never observed, and the frequencies are from 0 to approximately 600 000. The number of ktuples having these frequencies varies from 1 to approximately 300 000. One cause for this could be the small amount of data compared to the number of possible ktuples of length 12.

It was decided that ktuples of length 8 are too short to be of any interest. For k=12, approximately 250 000 different ktuples are only observed once. If the ktuples with frequencies < 20 are accumulated, the number reaches approximately 4 million. This number is too great based on several tests concerning the size of a python dictionary accomplished by the author. The decision was made to focus on ktuples of length 10 in further development.

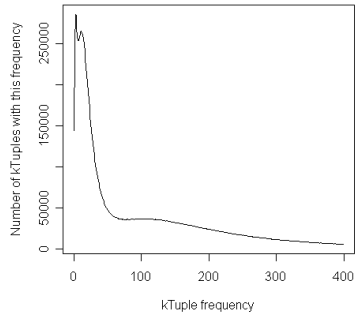
The frequencies of all the ktuples of length 10 were found for cow, and by assuming the hypothesis stated earlier, the next step was to extract the ones having a low frequency. These would later be used as the background for



(a)All the observed frequencies.



(b)A zoom in on the area having the most frequencies.



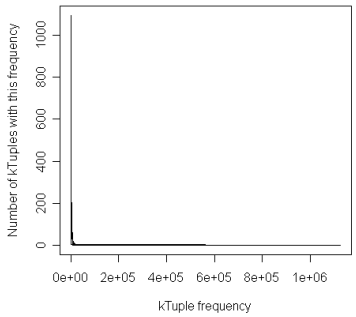
(c)Further zoom in on the lowest observed frequencies.

Figure 5.6: Plots showing the distribution of ktuples in the cow genome for k=12 in three levels.

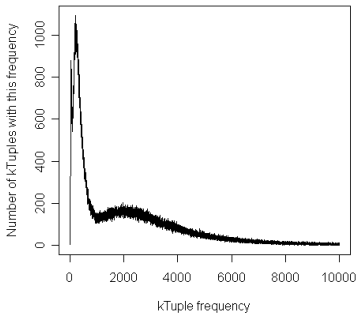
a search in the rice genome. The criteria defining what a low frequency is, was set to be a frequency of 20 as a maximum. A frequency of < 20 for a ktuple in the cow genome can be considered low when looking at the plots in Figure 5.7. Only a small proportion of the ktuples have this low frequency. A module meeting the condition mentioned above was implemented. The result was 1996 ktuples which met the criteria. These search-tuples were stored in a file.

5.5 The search in rice

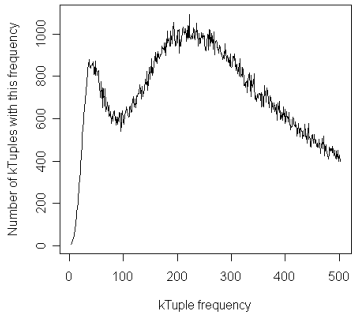
As described in the previous section, the ktuples with the lowest frequencies were found. The 1996 that met the criteria is called the search-tuples. The



(a)All the observed frequencies.



(b)A zoom in on the area having the most frequencies.



(c)Further zoom in on the lowest observed frequencies.

Figure 5.7: Plots showing the distribution of ktuples in the cow genome for k=10 in three levels.

next step was to investigate which of these are present in the rice genome.

Three different search methods for finding the search-tuples in the rice sequences was implemented. Two different properties of a ktuple is valuable, i) if it is present in rice, and if so ii) it's location in the genome. The first search method, called rice_search, looped through a sequence. For each subsequence of same length as the ktuple it checked if they were equal. Each time this test came true the file, sequence and position was stored in a dictionary with the ktuple as key. A test-run with a small file as input, revealed that this was a slow way to execute the task. A second approach called rice_search2, was implemented. Instead of a loop through the sequence to check for a ktuple, rice_search2 used a built-in method to find a substring in a string object. This one also turned out to be a slow solution. Both these approaches were altered to store only the number of times a ktuple was found in the sequences. This was done, since the first step to obtain is i).

For the rice_search method, the actual number of times a ktuple occur is found. For the second method on the other hand, a ktuple occurring several times in would only be registered one time. This is due to the nature of the built-in search method used in rice_search2. None of the two methods were performing as rapidly desired.

Hence a third search method called rice_search3 was implemented. In this one the searching method for fast text manipulation first used in HGT_nr2 was adopted. As described in the section for mxTextTools, the find-function in this package returns the position of the first occurrence. The result of this is the same as for rice_search2-method, that only one occurrence is found even if there are several. Due to this fact, it was decided to continue the idea to store counts rather than the positions in the dictionary at this point. The positions could be retrieved at a later time. A big difference compared to the two previous search methods was the high speed at which this implementation executed.

rice_search	rice_search2	rice_search3
2m40s	2m50s	1m43s

Table 5.2: The execution times for the three different search-methods for a search with all the 1996 ktuples in a file with size 133 kB.

As shown in Table 5.2, this third implementation improved execution time with approximately one minute compared to both the previous search methods. Considering that the size of the test file used was only a fraction of the total amount of data to be searched through, this improvement was very valuable.

An execution of the rice_search3 implementation in all the rice sequences was carried out. A scan of the search-result data showed that all 1996 ktuples involved in the search were found in the rice sequences. All of the search-tuples were found a minimum of 8 times and some even 14. Since each of the rice files contain only one sequence, this means that some of the ktuples occur in every rice chromosome. The total number of bases in the rice sequences are $372702846 = 3.7 \cdot 10^8$. If an even distribution is assumed, each ktuple of length 10 is expected to occur

$$\frac{3.7 \cdot 10^8}{4^{10}} \approx 352$$

times in the rice genome. As with the cow genome, it is safe to assume

that the rice genome does not have an even distribution. Hence some of the ktuples might occur several thousand times.

The fundamental reasoning in this thesis was to use the ktuples present in the dog genome to filter out the ktuples which were found in both cow and rice, before expanding and aligning. As can be concluded based on the previous calculations, all of the ktuples will probably be present in the dog sequences as well. Thus if the filtering of ktuples was to be carried out, none would be left to work with. This step was dropped, and other measures was taken.

5.6 Expanding the tuples

The biological value of long DNA sequences are more interesting than that of short DNA sequences. To achieve more significant biological results, the ktuples should be expanded. The next step was to find all occurrences of the 1996 search-tuples, both in the cow and the rice genome and expand them. The ktuples were expanded by retrieving the 210 bp long subsequence consisting of 100 bases in front of the ktuple to 100 bases succeeding it. This will be referred to as an expansion or extension. Figure 5.8 illustrates this expansion of the ktuples.

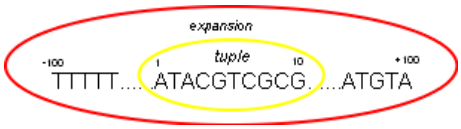


Figure 5.8: How the ktuples are to be extended.

These expansions would afterwards be subject to alignments. For a ktuple, each expansion in cow was to be aligned with every expansion in rice. A moment is taken to reflect on the number of expansions and the enormous amount of alignments that comes with them. The number of alignments to be carried out is $n \cdot m$ for each search-tuple. n is the number of occurrences in cow and m is the number in rice. An example follows. Assume that all 1996 search-tuples occur only 14 times each in rice and 10 times each in cow. Both these numbers are understatements, still they give a total of $1996 \cdot 14 \cdot 10 = 279440$ alignments. The probability that several of the ktuples occur hundreds of times in both genomes is high. This may result in several million alignments.

5.6.1 Reduction of the search-tuples

To reduce the number of alignments, the criteria for which of the search-tuples that are to be extended was strengthened as to only concern those with frequencies ≤ 10. Figure 5.9 shows a plot where the frequency limit of 10 is drawn. 248 ktuples had a frequency ≤ 10. A module for extract-

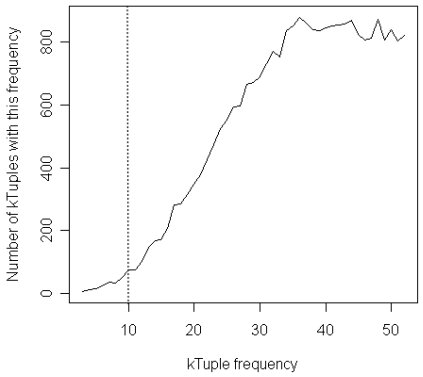


Figure 5.9: A plot where the limits for which ktuples are to be extended are shown.

ing all expansions in both organisms for all search-tuples having frequencies matching the reinforced criteria was implemented. After excluding from the expansions those which include 'N' as a base, all expansions were saved in two separate files. One contained those retrieved from the cow sequences and the other those from rice. Table 5.3 shows some properties of these search-tuples. As could be calculated from the given average values in Table 5.3,

	Occurrences	Average occurrence	Min freq	Max freq
Cow	1989	8	3	10
Rice	30979	125	18	580

Table 5.3: Informational values for the 248 search-tuples.

this gave an average of 1000 alignments per ktuple. With 248 ktuples this still was a large number.

To get an acceptable number of alignments, two approaches seemed appropriate:

1. Continue with the ktuples having the lowest frequency in cow which

was 3, and not consider the frequency in rice at all. This approach focuses on the theory that a low frequency in cow may imply that the ktuples originally came from rice, hence their frequency is this low. It assumes nothing about whether the frequency of a ktuple in rice affects a conceivable transfer to cow. This method will be referred to simply as **LP** - low frequency in the primary organism.

2. Continue with ktuples of the two lowest frequencies (3 and 4) but only consider those that in addition have a high frequency in rice. This approach implies that a high frequency in rice enhances the probability for a transfer to cow. This method will be referred to as **LPHF** - low frequency in the primary organism and high frequency in the feed.

Now that the number of ktuples, and thereby expansions, lead to an acceptable number of alignments the next step was to perform the aligning.

5.7 Alignments

A module that performs the alignments was implemented. Each of the two methods for selecting ktuples was executed, and for every ktuple meeting the criteria, the expansions in cow were aligned with the corresponding extensions in rice. A more detailed description on how the alignments are done follows.

As stated in the background section, the Biopython project includes a package called Bio.pairwise2, which was used for performing the alignments. The variant for local alignment was used, since it is assumed to be a relatively small amount of similar segments in sequences from cow and rice. An affine gap-penalty is applied, and a scoring matrix storing the score for any pair of characters is given. A call with

```
pairwise2.align.localds(cow_seq,rice_seq,match_dict,open,extend,one=True)
```

gives a highest scoring local alignment of two sequences. The scoring matrix used is that shown in Table 5.4. It is inspired by the Kimura two-parameter model, and gives a higher score to transitions (A-G,C-T) over transversions (A-C,A-T,G-T,C-G) (Kimura, 1980). The gap-open penalty is 5.0 and the gap-extend penalty is 3.0.

	A	C	G	T
A	5.0	-3.0	-2.0	-3.0
C		5.0	-3.0	-2.0
G			5.0	-3.0
T				5.0

Table 5.4: The scoring matrix used for the alignments.

Alignment nr	LP score
1	392
2	377
3	358
4	351
5	347
6	346
7	345
8	341
9	338
10	337

(a)

Alignment nr	LPHF score
1	405
2	395
3	395
4	393
5	392
6	391
7	390
8	389
9	387
10	387

(b)

Table 5.5: The resulting scores for the top 10 alignments for the two different ktuple selecting methods, Low frequency in the Primary organism **(a)**, and Low frequency in the Primary organism and High frequency in Feed **(b)**. The alignments are different for the two methods.

For the two methods, for each ktuple that meets the criteria, all expansions of that ktuple in cow was aligned with all expansions of it in rice. Each of these produced an alignment. All alignments for all ktuples were found, and the 10 highest scoring alignments for the two methods were stored in files.

As shown in Table 5.5, the alignments based on the LPHF method achieved higher scores than those from the LP method, except for the nr.1 highest scoring alignment. For the LP method the highest score was 392, while LPHF had a top-score of 405. The actual scores of the alignments are not really important, since they are mere used as a measure to distinguish the two methods. But what was important and that can be interpreted from the observations in the Table 5.5, was that the LPHF approach seemed to find sequences more similar. The highest scoring alignment for the LPFH methods is shown in Figure 5.10. The top 10 alignments found using the LPFH method were used further in the quest.

```
>TCGTCGAACG | nr.1
GCCGCCGCTC-TATGACC---TCCGCGCCGCCG-CCGCCGCCGCTCGTCGGCCCACGCGGCAA-G
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
GTCGTCG-TCGTCCCACCCCATCC-CGCCGCCGAACG-CGCCGCCGGGGGGCCC-CA-GG-AACC

TGGGC--GGGACCCGGGCGG-GCTTCGGACTCCCGCGCCCGGC-C-TCGTCGAACGGCGGCCACG
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CGAGCTTCTG-CCCCGGCGGCGC-GAGGA---ACGGGGGCGGCGCGTCGTCGAACG--GGT--TG

G-GGAGGCCGAAGGACTCCTGGA--GGCGCGGGC-CCAG--GGCG-CG-AGGCGGCGTCGCAGTA
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
GAGGAGG-CGAAG---TCGTGGAACGGCGGGAGCGCC-GCCGTCGTCGTGGGCGGCG--GC-G--

GCGCG-CGG-GGACGACTG--CCTCCAGCCAAGAGGCC---
|||||||||||||||||||||||||||||||||||||||
GCGGGACGGAGGACGACGGGAAGGCC-GCCTGGAAGTCGAA
      Score = 405

| = identical nucleotide
| or | = included in the score.
```

Figure 5.10: The highest scoring alignment achieved by using the LPHF method to select the ktuples. The ktuple used as basis for the expansion is written in blue.

5.8 Searches using BLAST

Involved in the top ten alignments were three different cow and ten different rice sequences. The three 210 bp long cow sequences are

```
> nr.1
GCCGCCGCTCTATGACCTCCGCGCCGCCGCCGCCGCCGCTCGTCGGCCCACGCGGCAAGT
GGGCGGGACCCGGGCGGGCTTCGGACTCCCGCGCCCGGCCTCGTCGAACGGCGGCCACGG
GGAGGCCGAAGGACTCCTGGAGGCGCGGGCCCAGGGCGCGAGGCGGCGTCGCAGTAGCGC
GCGGGGACGACTGCCTCCAGCCAAGAGGCC
> nr.2
CCCACGGCGGGATGCGGCCTCTCCAAGTGCTGGGCCGGGTCAGCGCGATGCCCGCCTGGG
TCCCGGCGCGGGGAAAGAGCGCGGAGGCCCAGGCCCCTCCACGATCGCGATACCGAGGCC
CGGGAAGAGCAGGACCCCGATGGGCCCCGCGGCAGCGGCGCGCGGCGCCGAGCCTGGACC
CCGCTTGCAGCCCCGCCCGGGTCCCCCTCC
> nr.3
TTTCCAGGAAGGGAGCGCCCAGTGCGCTCCCGGGCCGGGTGGAGGCTAAGGCAGAGGGAG
ACTTCGACTCTATCGCTGTTTGCTCTGCAGCCGGGCCAGATCGACGCGTACGCGTGCCTT
GGCGGTCCTGCGGTCACCCTCCCGAGTCCTGGCCGTGCAGTCGATCGGGCCGTCGGCGAT
GTTTTATTTCCACTGTCCACCACAGCTAGA
```

These will in the following be referred to as nr.1, nr.2 and nr.3 respectively.

Each of the sequences involved in the top ten alignments were used as query sequences in a manually accomplished BLAST search. A BLAST search can help establish the true origin of the sequences from the two organisms. The BLASTN search for short, nearly exact matches was used. The settings were nr as the nucleotide database, no filters on, an expect value of 1000 and a word size of 7.

The BLAST results for all ten rice sequences verified beyond any doubt that they are of rice origin. The BLAST results for the cow sequences is presented next. The BLAST outputs and the corresponding distance trees for each of the three sequences are given in Figures 5.11 to 5.16.

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
NM_001046242.1	Bos taurus telomerase reverse transcriptase (TERT), mRNA	416	416	100%	1e-113	100%
DQ464899.1	Bos taurus telomerase reverse transcriptase transcript variant 2 (TERT) mRNA, complete cds, alternatively spliced	416	416	100%	1e-113	100%
DQ464898.1	Bos taurus telomerase reverse transcriptase transcript variant 1 (TERT) mRNA, complete cds, alternatively spliced	416	416	100%	1e-113	100%
DQ399842.2	Bos taurus telomerase reverse transcriptase (TERT) gene, complete cds	416	416	100%	1e-113	100%
CT834590.1	Oryza sativa (indica cultivar-group) cDNA clone:OSIGCSA060L15, full insert sequence	52.0	120	12%	8e-04	100%
NM_001063238.1	Oryza sativa (japonica cultivar-group) Os06g0132600 (Os06g0132600) mRNA, complete cds	52.0	52.0	12%	8e-04	100%
AF488413.1	Oryza sativa chromosome 6 BAC 134P10, complete sequence	52.0	222	12%	8e-04	100%
AP008212.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 6	52.0	6.985e+04	56%	8e-04	100%
AP002542.2	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 6, PAC clone:P0679C08	52.0	257	12%	8e-04	100%
AK074010.1	Oryza sativa (japonica cultivar-group) cDNA clone:J033076A22, full insert sequence	52.0	52.0	12%	8e-04	100%
AK061226.1	Oryza sativa (japonica cultivar-group) cDNA clone:006-210-H12, full insert sequence	52.0	52.0	12%	8e-04	100%
CT836040.1	Oryza sativa (indica cultivar-group) cDNA clone:OSIGCSN088I10, full insert sequence	50.1	86.2	11%	0.003	100%
NM_001065811.1	Oryza sativa (japonica cultivar-group) Os07g0243200 (Os07g0243200) mRNA, complete cds	50.1	50.1	11%	0.003	100%
NM_001052768.1	Oryza sativa (japonica cultivar-group) Os02g0202700 (Os02g0202700) mRNA, complete cds	50.1	50.1	11%	0.003	100%
BC105925.1	Homo sapiens cytoplasmic polyadenylation element binding protein 2, mRNA (cDNA clone MGC:125377 IMAGE:40008495), complete c	50.1	84.2	11%	0.003	100%
BC103940.1	Homo sapiens cytoplasmic polyadenylation element binding protein 2, mRNA (cDNA clone MGC:119575 IMAGE:40008489), complete c	50.1	84.2	11%	0.003	100%
NM_182485.1	Homo sapiens cytoplasmic polyadenylation element binding protein 2 (CPEB2), transcript variant B, mRNA	50.1	84.2	11%	0.003	100%
NM_182646.1	Homo sapiens cytoplasmic polyadenylation element binding protein 2 (CPEB2), transcript variant A, mRNA	50.1	84.2	11%	0.003	100%
AY255519.1	Homo sapiens cytoplasmic polyadenylation element binding protein CPEB2b mRNA, complete cds	50.1	84.2	11%	0.003	100%
AY247744.1	Homo sapiens cytoplasmic polyadenylation element binding protein CPEB2 mRNA, complete cds	50.1	84.2	11%	0.003	100%
AP008213.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 7	50.1	5.803e+04	52%	0.003	100%
AP008208.1	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 2	50.1	7.988e+04	77%	0.003	100%
AC105289.4	Homo sapiens BAC clone RP11-629A7 from 4, complete sequence	50.1	283	14%	0.003	100%
AP003754.5	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 7, BAC clone:OJ1341_A08	50.1	456	11%	0.003	100%
AP004045.3	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 2, BAC clone:OJ1135_F06	50.1	259	12%	0.003	100%
AP004382.2	Oryza sativa (japonica cultivar-group) genomic DNA, chromosome 7, PAC clone:P0418E08	50.1	349	11%	0.003	100%
AK121036.1	Oryza sativa (japonica cultivar-group) cDNA clone:J023050O10, full insert sequence	50.1	50.1	11%	0.003	100%
BC103941.1	Homo sapiens cytoplasmic polyadenylation element binding protein 2, mRNA (cDNA clone MGC:119576 IMAGE:40008490), complete c	50.1	120	11%	0.003	100%
CP000554.1	Prochlorococcus marinus str. MIT 9303, complete genome	48.1	148	11%	0.013	100%
NM_011381.3	Mus musculus sine oculis-related homeobox 3 homolog (Drosophila) (Six3), mRNA	48.1	48.1	11%	0.013	100%
CT831722.1	Oryza sativa (indica cultivar-group) cDNA clone:OSIGCPI015H11, full insert sequence	48.1	120	11%	0.013	100%
AC187512.2	Gallus gallus BAC clone CH261-111G7 from chromosome z, complete sequence	48.1	222	11%	0.013	100%
NM_001073351.1	Oryza sativa (japonica cultivar-group) Os12g0500700 (Os12g0500700) mRNA, complete cds	48.1	48.1	11%	0.013	100%
NM_001068686.1	Oryza sativa (japonica cultivar-group) Os08g0499900 (Os08g0499900) mRNA, complete cds	48.1	80.3	11%	0.013	100%
NM_001064785.1	Oryza sativa (japonica cultivar-group) Os06g0655500 (Os06g0655500) mRNA, complete cds	48.1	84.2	17%	0.013	93%
NM_001056271.1	Oryza sativa (japonica cultivar-group) Os03g0280000 (Os03g0280000) mRNA, complete cds	48.1	152	11%	0.013	100%
NM_001052887.1	Oryza sativa (japonica cultivar-group) Os02g0227400 (Os02g0227400) mRNA, complete cds	48.1	48.1	11%	0.013	100%
NM_001051287.1	Oryza sativa (japonica cultivar-group) Os01g0839300 (Os01g0839300) mRNA, complete cds	48.1	120	11%	0.013	100%

Figure 5.11: The BLAST result for cow sequence nr.1. The result included several other sequences in addition to those shown, but they are cut out due to the corresponding high E-value.

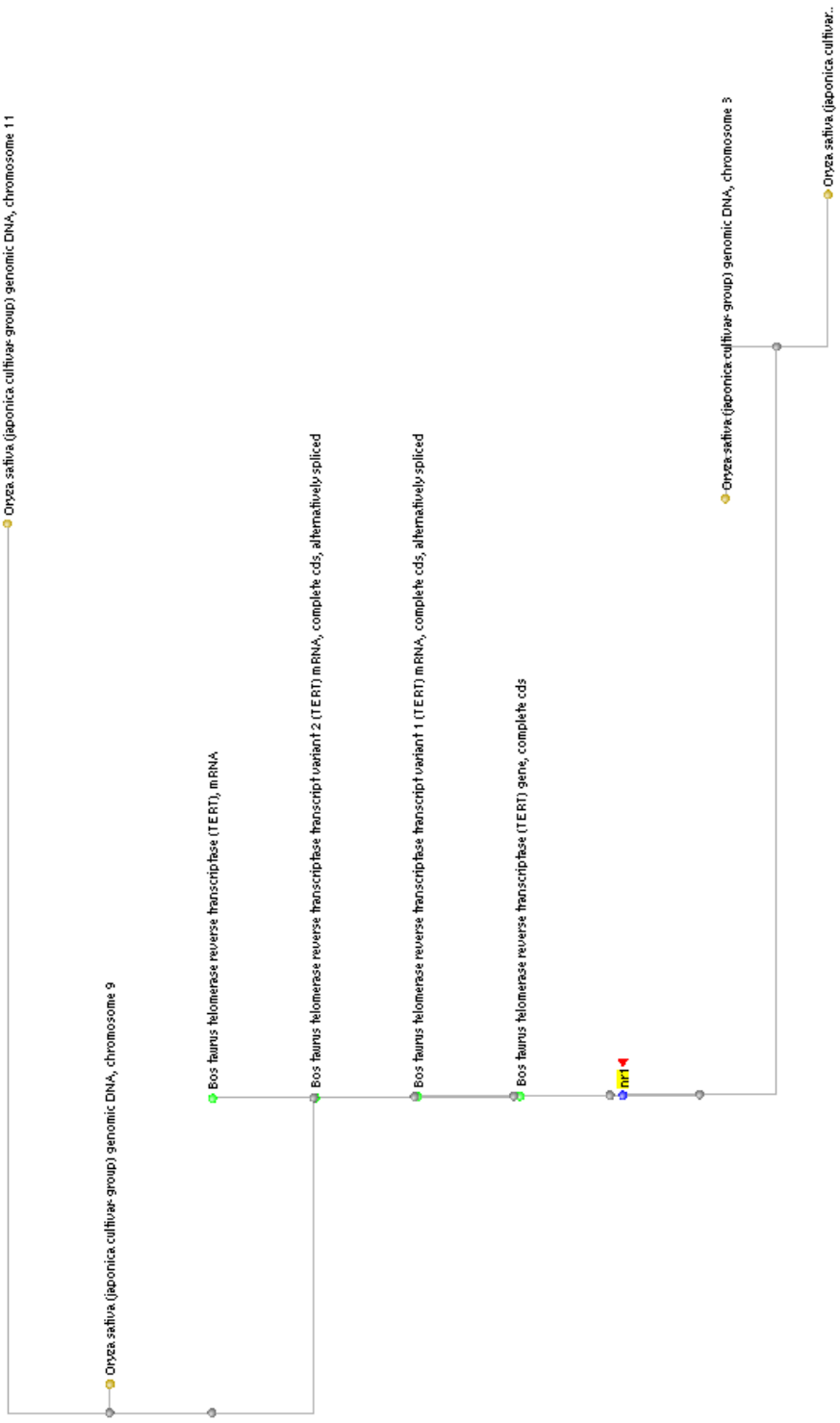


Figure 5.12: The distance tree generated from the BLAST output for cow sequence nr.1.

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
CP000578.1	Rhodobacter sphaeroides ATCC 17029 chromosome 2, complete sequence	46.1	148	14%	0.050	100%
CP000570.1	Burkholderia pseudomallei 668 chromosome I, complete sequence	42.1	1052	30%	0.79	100%
CP000577.1	Rhodobacter sphaeroides ATCC 17029 chromosome 1, complete sequence	42.1	249	27%	0.79	100%
XR_028795.1	PREDICTED: Bos taurus hypothetical LOC538483 (LOC538483), mRNA	42.1	42.1	10%	0.79	100%
CP000143.1	Rhodobacter sphaeroides 2.4.1 chromosome 1, complete genome	42.1	249	27%	0.79	100%
CP000548.1	Burkholderia mallei NCTC 10247 chromosome II, complete sequence	40.1	887	30%	3.1	100%
CP000572.1	Burkholderia pseudomallei 1106a chromosome I, complete sequence	40.1	1050	33%	3.1	100%
CP000546.1	Burkholderia mallei NCTC 10229 chromosome II, complete sequence	40.1	887	30%	3.1	100%
AC195098.4	Rhesus Macaque BAC CH250-177N5 () complete sequence	40.1	72.4	9%	3.1	100%
CP000526.1	Burkholderia mallei SAVP1 chromosome II, complete sequence	40.1	887	30%	3.1	100%
CP000459.1	Burkholderia cenocepacia HI2424 chromosome 2, complete genome	40.1	332	22%	3.1	100%
CP000458.1	Burkholderia cenocepacia HI2424 chromosome 1, complete genome	40.1	466	15%	3.1	100%
NM_001063780.1	Oryza sativa (japonica cultivar-group) Os06g0237000 (Os06g0237000) mRNA, complete cds	40.1	72.4	11%	3.1	100%
CP000440.1	Burkholderia cepacia AMMD chromosome 1, complete sequence	40.1	301	32%	3.1	100%
AM236080.1	Rhizobium leguminosarum bv. viciae chromosome complete genome, strain 3841	40.1	271	36%	3.1	100%
XM_001164898.1	PREDICTED: Pan troglodytes apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3D, transcript variant 6 (APOBEC	40.1	40.1	9%	3.1	100%
XM_001164934.1	PREDICTED: Pan troglodytes apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3D, transcript variant 7 (APOBEC	40.1	40.1	9%	3.1	100%
AC187604.3	Pan troglodytes BAC clone CH251-739K24 from chromosome 7, complete sequence	40.1	72.4	9%	3.1	100%
NM_015879.2	Homo sapiens ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 3 (ST8SIA3), mRNA	40.1	40.1	9%	3.1	100%
XM_001101953.1	PREDICTED: Macaca mulatta aldehyde dehydrogenase 3 family, member A1, transcript variant 1 (ALDH3A1), mRNA	40.1	40.1	9%	3.1	100%
XM_001102132.1	PREDICTED: Macaca mulatta aldehyde dehydrogenase 3 family, member A1, transcript variant 2 (ALDH3A1), mRNA	40.1	40.1	9%	3.1	100%
CP000379.1	Burkholderia cenocepacia AU 1054 chromosome 2, complete sequence	40.1	299	22%	3.1	100%
CP000378.1	Burkholderia cenocepacia AU 1054 chromosome 1, complete sequence	40.1	400	23%	3.1	100%
AC004633.1	Homo sapiens chromosome 5, P1 clone 41e13 (LBNL H151), complete sequence	40.1	40.1	9%	3.1	100%
CP000010.1	Burkholderia mallei ATCC 23344 chromosome 1, complete sequence	40.1	951	30%	3.1	100%
CP000124.1	Burkholderia pseudomallei 1710b chromosome I, complete sequence	40.1	1050	32%	3.1	100%
AC134930.2	Oryza sativa (japonica cultivar-group) chromosome 5 clone OSJNBb0042J17, complete sequence	40.1	40.1	11%	3.1	95%
AC004125.1	Homo sapiens Chromosome 16 BAC clone CIT987SK-625P11, complete sequence	40.1	72.4	9%	3.1	100%
AC015909.14	Homo sapiens chromosome 17, clone RP11-893F2, complete sequence	40.1	72.4	9%	3.1	100%
AC010971.3	Homo sapiens BAC clone RP11-122G11 from 7, complete sequence	40.1	72.4	9%	3.1	100%
BC007954.2	Homo sapiens cDNA clone IMAGE:4302702, partial cds	40.1	40.1	9%	3.1	100%
AC090208.4	Homo sapiens chromosome 18, clone RP11-660C14, complete sequence	40.1	40.1	9%	3.1	100%
AC079326.8	Homo sapiens chromosome 11, clone RP11-233I10, complete sequence	40.1	40.1	9%	3.1	100%
AC021607.9	Homo sapiens chromosome 18, clone RP11-233O10, complete sequence	40.1	40.1	9%	3.1	100%
AC093928.2	Homo sapiens chromosome 3 clone RP11-241K3, complete sequence	40.1	40.1	9%	3.1	100%
AC116353.2	Homo sapiens chromosome 5 clone RP11-515C16, complete sequence	40.1	40.1	9%	3.1	100%
AC090340.6	Homo sapiens chromosome 18, clone RP11-714M23, complete sequence	40.1	40.1	9%	3.1	100%
AC091956.3	Homo sapiens chromosome 5 clone RP11-430F16, complete sequence	40.1	40.1	9%	3.1	100%

Figure 5.13: The BLAST result for cow sequence nr.2. The result included several other sequences in addition to those shown, but they are cut out due to the corresponding high E-value.

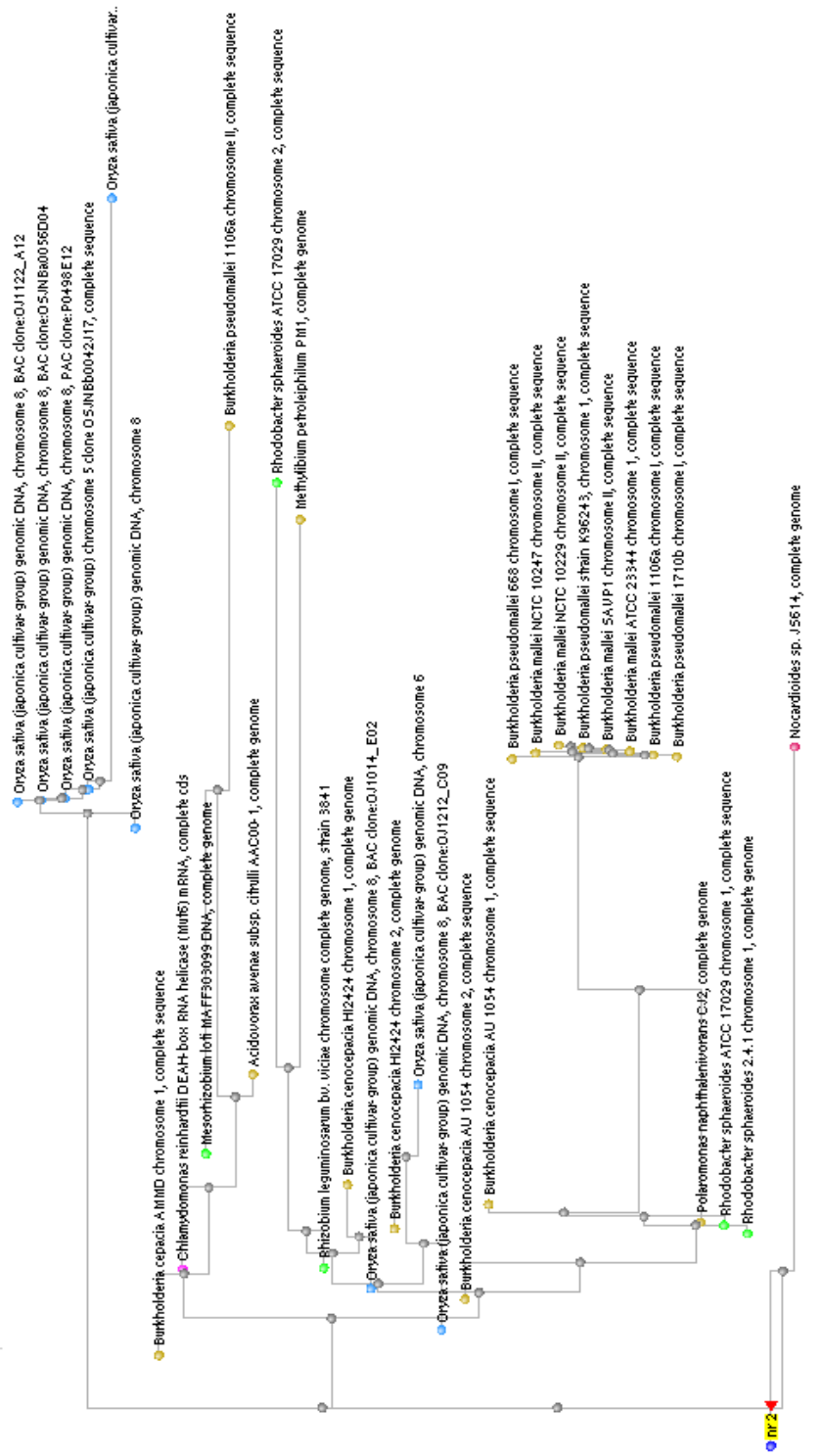


Figure 5.14: The distance tree generated from the BLAST output for cow sequence nr.2.

Accession	Description	Max score	Total score	Query coverage	E value	Max ident
BC132613.1	Mus musculus paired related homeobox protein-like 1, mRNA (cDNA clone MGC:164244 IMAGE:40130890), complete cds	73.8	73.8	19%	2e-10	97%
NM_001001796.1	Mus musculus paired related homeobox protein-like 1 (Prrxl1), mRNA >dbj|AK039633.1| Mus musculus adult male spinal cord cDNA, RIKEN	73.8	73.8	19%	2e-10	97%
CT009541.12	Mouse DNA sequence from clone RP23-56C21 on chromosome 14, complete sequence	73.8	110	33%	2e-10	97%
XM_843626.1	PREDICTED: Canis familiaris similar to paired related homeobox protein-like 1 (LOC607056), mRNA	71.9	142	54%	9e-10	95%
NM_145767.1	Rattus norvegicus paired related homeobox protein-like 1 (Prrxl1), mRNA >gb|U29174.1|RNU29174 Rattus norvegicus paired-like homeod	65.9	65.9	19%	5e-08	95%
AC027674.10	Homo sapiens chromosome 10 clone RP11-255M1, complete sequence	58.0	106	38%	1e-05	94%
AL138760.14	Human DNA sequence from clone RP11-123B3 on chromosome 10, complete sequence	58.0	106	38%	1e-05	94%
CP000509.1	Nocardioides sp. JS614, complete genome	48.1	285	32%	0.013	100%
AL359636.17	Human DNA sequence from clone RP11-542K23 on chromosome 9, complete sequence	44.1	44.1	10%	0.20	100%
CP000555.1	Methylibium petroleiphilum PM1, complete genome	42.1	138	15%	0.79	100%
BX572594.1	Rhodopseudomonas palustris CGA009 complete genome; segment 2/16	42.1	42.1	10%	0.79	100%
XM_001365976.1	PREDICTED: Monodelphis domestica similar to paired-like homeodomain transcription factor (LOC100014678), mRNA	40.1	40.1	9%	3.1	100%
XM_585577.3	PREDICTED: Bos taurus similar to transcription factor Tbx4 (LOC508750), mRNA	40.1	40.1	9%	3.1	100%
XM_001110382.1	PREDICTED: Macaca mulatta similar to T-box 4 (LOC712702), mRNA	40.1	40.1	9%	3.1	100%
AC162176.2	Mus musculus BAC clone RP24-446P15 from chromosome 9, complete sequence	40.1	40.1	9%	3.1	100%
AC123850.3	Mus musculus BAC clone RP23-319K22 from chromosome 5, complete sequence	40.1	40.1	9%	3.1	100%
AE011839.1	Xanthomonas axonopodis pv. citri str. 306, section 217 of 469 of the complete genome	40.1	40.1	9%	3.1	100%
AE012377.1	Xanthomonas campestris pv. campestris str. ATCC 33913, section 285 of 460 of the complete genome	40.1	40.1	11%	3.1	95%
AP006627.1	Bacillus clausii KSM-K16 DNA, complete genome	40.1	72.4	19%	3.1	100%
CP000050.1	Xanthomonas campestris pv. campestris str. 8004, complete genome	40.1	168	25%	3.1	100%
BA000040.2	Bradyrhizobium japonicum USDA 110 DNA, complete genome	40.1	172	14%	3.1	100%
AC004057.1	Homo sapiens chromosome 4 clone B240N9 map 4q25, complete sequence	40.1	40.1	11%	3.1	95%
CP000573.1	Burkholderia pseudomallei 1106a chromosome II, complete sequence	38.2	172	29%	12	100%
CP000571.1	Burkholderia pseudomallei 668 chromosome II, complete sequence	38.2	172	29%	12	100%
CP000493.1	Hyperthermus butylicus DSM 5456, complete genome	38.2	38.2	9%	12	100%
CP000512.1	Acidovorax avenae subsp. citrulli AAC00-1, complete genome	38.2	172	27%	12	100%
AM236085.1	Rhizobium leguminosarum bv. viciae plasmid pRL11 complete genome, strain 3841	38.2	38.2	9%	12	100%
CP000353.1	Ralstonia metallidurans CH34 megaplasmid, complete sequence	38.2	102	16%	12	100%
AC134236.4	Oryza sativa (japonica cultivar-group) chromosome 3 clone OSJNBa0056G13, complete sequence	38.2	38.2	9%	12	100%
AC140182.2	Mus musculus BAC clone RP24-349I16 from chromosome 8, complete sequence	38.2	38.2	10%	12	95%
CP000125.1	Burkholderia pseudomallei 1710b chromosome II, complete sequence	38.2	172	29%	12	100%
AC166779.3	Mus musculus chromosome 18, clone RP23-399N10, complete sequence	38.2	38.2	9%	12	100%
AC157910.4	Mus musculus chromosome 18, clone RP24-372A6, complete sequence	38.2	38.2	9%	12	100%
AC126421.12	Mus musculus chromosome 1, clone RP23-167H24, complete sequence	38.2	38.2	9%	12	100%
AC102358.5	Mus musculus chromosome 8, clone RP23-451H9, complete sequence	38.2	38.2	10%	12	95%
AC125039.4	Mus musculus BAC clone RP23-335C3 from chromosome 6, complete sequence	38.2	38.2	9%	12	100%
AC124202.3	Mus musculus BAC clone RP23-337F21 from chromosome 14, complete sequence	38.2	38.2	9%	12	100%

Figure 5.15: The BLAST result for cow sequence nr.3. The result included several other sequences in addition to these, but they are cut out due to the corresponding high E-value.

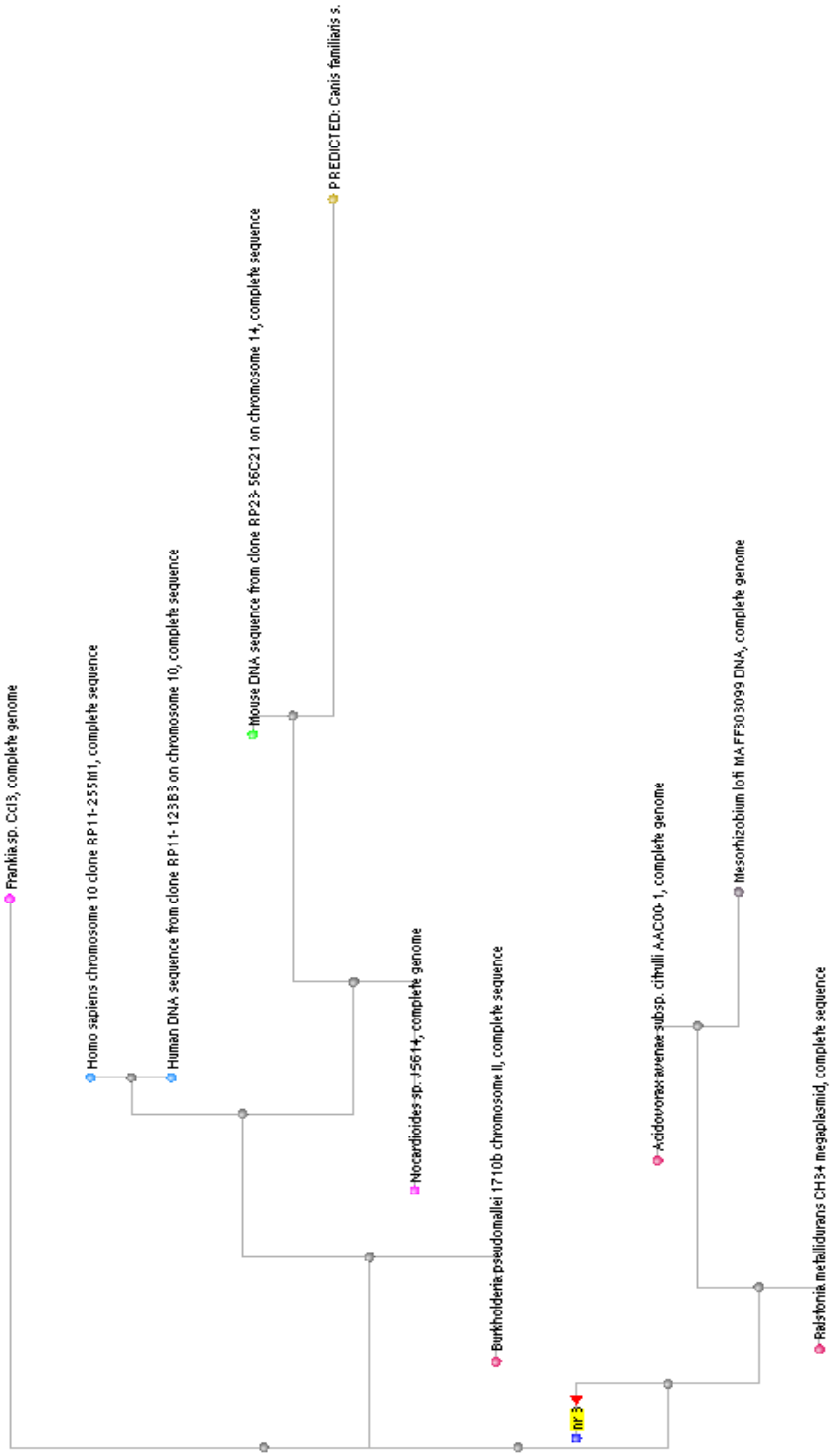


Figure 5.16: The distance tree generated from the BLAST output for cow sequence nr.3.

5.9 Interpretation of the results

The nr.1 search sequence is found in the following cow-sequence:

```
gi|76646740|ref|NW\_929937.1|Bt20\_WGA3334\_2 Bos taurus
chromosome 20 genomic contig
```

The best matches for the nr.1 sequence were highly significant matches with *bos taurus* having a 100% query coverage. This confirmed that the sequence origins from cow. Other significant matches were achieved with *oryza sativa*. Most of these had a query coverage of 12% and 100% identity, which means that a segment of length 26 in the search sequence is completely matched by rice sequences. One of these hits with rice is shown next:

```
>gb|AF488413.1|Oryza sativa chromosome 6 BAC 134P10,
complete sequence
Length=136866

 Score = 52.0 bits (26),  Expect = 8e-04
 Identities = 26/26 (100\%), Gaps = 0/26 (0\%)
 Strand=Plus/Plus

Query  17     CTCCGCGCCGCCGCCGCCGCCGCTCG  42
              ||||||||||||||||||||||||||
Sbjct  80018  CTCCGCGCCGCCGCCGCCGCCGCTCG  80043
```

A closer look at all the matches from rice with an E-value less than 0.01 revealed that they were all in the same area in nr.1. These hits were located in positions 14 to 47, with varying sizes of 24 to 26 bp. A further investigation showed that they are all microsatellite DNA. Microsatellite DNA is as described by McDonald (2007) a sequence of 2 to 5 nucleotides that are repeated. The sequence `CGC` is the microsatellite in this case. Microsatellites may occur thousands of times in the genome, hence they cannot be used as proof that a horizontal transfer has taken place.

For both the nr.2 and the nr.3 sequence, no significant match with *bos taurus* sequences were achieved. This was an unexpected result. It is reasonable to assume that a short sequence extracted from a larger cow sequence origins

from cow. To be a hundred percent certain that these two sequences were extracted from the cow files downloaded, a test was performed which confirmed their presence. All DNA sequences in a genome database have an accession code. The *bos taurus* sequences in which the nr.2 and nr.3 sequences are found have accession codes NW_936280.1 and NW_928695.1. A search on the NCBI homepage (NCBI, 2006*a*) for these accession codes in march 2007 received the following answer: "This record was removed as a result of standard genome annotation processing." The cow files used in this project are now outdated. The *bos taurus* genome sequences have been modified and the latest draft was published in January this year. This is the most likely reason why the nr.2 and nr.3 sequences are not significant homologues with *bos taurus*.

5.10 kTuple Frequency Based method illustrated

In the previous sections the steps in the method called the kTuple Frequency Based are method presented. As a summary, Figure 5.17 illustrates these steps.

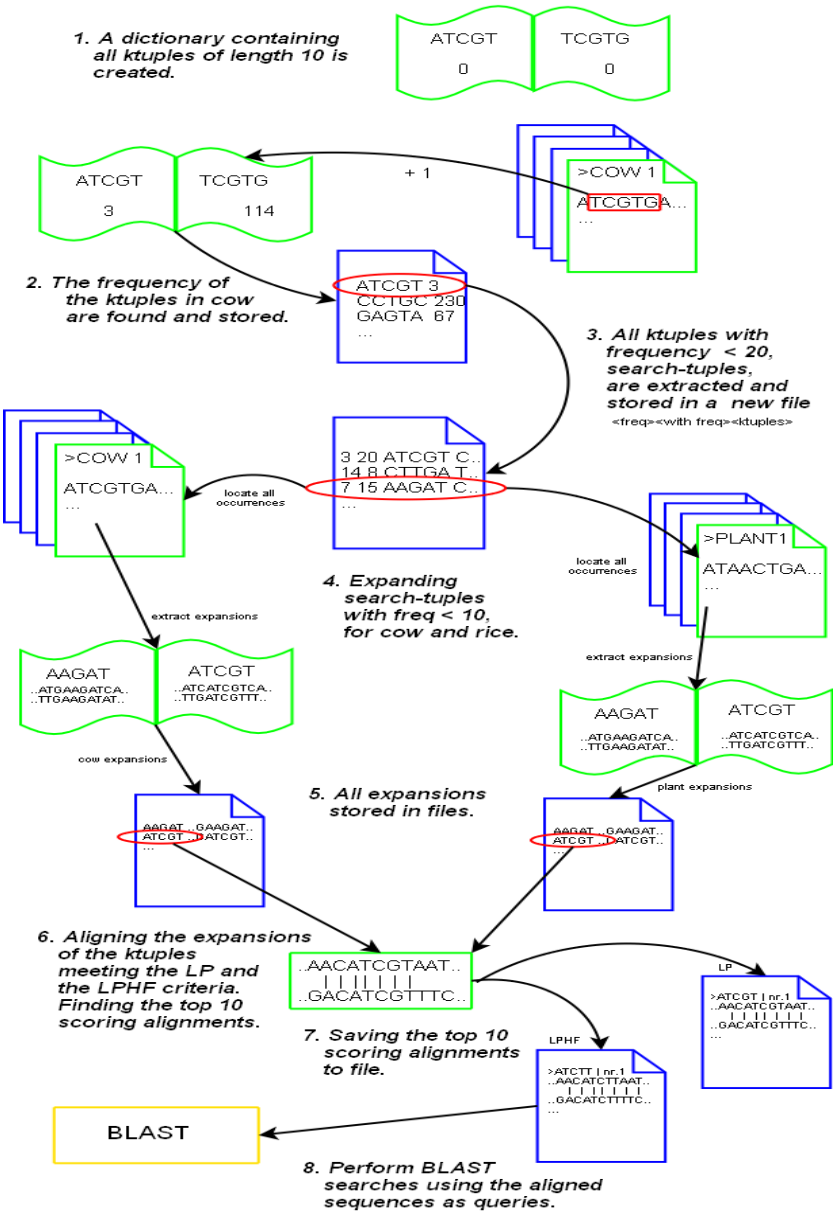


Figure 5.17: Illustration of the kTuple Frequency Based method. The different steps are shown.

5.11 Simulation

The procedure proposed in this thesis, the kTuple Frequency Based method, is a new way of approaching the disclosure of possible horizontal transfer eventz. It is hard to know how solid the procedure is. One way to get an estimate of this is to perform a simulation.

The principal thought in the kTuple Frequency Based method is that if a segment of DNA has been horizontally transferred into an organism, this segment should occur less often in the genome than subsequences originally in this organism. The 1996 search-tuples of length 10 in the approach are all chosen on basis of their low frequency in cow, the primary organism.

Assuming that a HGT from rice to cow has occurred, how good is the proposed algorithm at discovering such an event? To simulate that HGT events from rice to cow has occured, one thousand sequences of a certain length are randomly picked from the rice genome. The number of times at least one of the 1996 search-tuples is present in these random rice sequences is calculated. This simulation does not take into consideration any mutations. This is a simplification of real life, and gives inaccurate results.

Length	% hits
1000	31.1
500	19.6
250	9.5
100	5.3
50	1.7
10	0.2

Table 5.6: The simulated probability for finding at least one search-tuple in a random sequence of known length with the kTuple Frequency Based method

As shown in Table 5.6, there is a 31% chance that a HGT event of a 1000 bp long sequence will be discovered by this approach, if all the 1996 search-tuples is used. The discovery-rate decreases with decreasing size of a transferred sequence. This is only natural as it is harder to find a needle than a shovel in a haystack.

The execution performed in this thesis is a mini-version of what is simulated, due to the fact that the number of search-tuples have dropped to 4. A consequence of this reduction is that the chances for discovering any HGT

events are minimal.

Chapter 6

Discussion and conclusions

6.1 Discussion

6.1.1 The abandoned methods

The basis algorithm presented in chapter 6 focuses on ktuples of length 30. This gives a large number of different ktuples to handle, $1.15 \cdot 10^{18}$ to be exact. The implementations of the HGT_nr1, HGT_nr2 and HGT_nr2 version 2 methods, led to the conclusion that the enormous amount of data involved when considering ktuples of length 30 was impossible to handle with the applied implementations. In retrospect, these conclusions were perhaps a bit hasty. As the understanding of the material has matured, it has become clear that alternative angles of implementing the basis algorithm might be successful.

The HGT_nr1 method keeps all the ktuples of length 30 in a dictionary. Other storing units such as lists were also tried, but none of them were capable of holding all ktuples present in the cow genome. One idea for a change which could have been tested is to divide the ktuples into groups. These groups should be small enough to be kept in a dictionary in the internal memory. One way to divide the ktuples could be based on their base-composition. The focus should be on one group at a time. When not in use these could be saved on the disc. Another improvement could have been to use the method from SSAHA (Ning et al., 2001) to convert from a nucleotide sequence to a digit. All sequences of a certain length gets an unambiguous number, which

could be used as index in a list. Because of the great amount of different ktuples, there would have to be several lists, each with a fixed number of positions. If instead of the number of times a ktuple is present the presence or absence of the ktuple was determined, each list would consist only of 0 or 1 in each position, instead of a possibly great number in addition to a 30 character string.

Both versions of the HGT_nr2 method involved a huge amount of I/O-operations, while just a small piece of data was kept in memory. One improvement could be to store the ktuples in memory, and not perform the search for the ktuples before a limit value for a number of ktuples was reached. Searching for all of the stored ktuples in the files at the same time, instead of one by one, would be more time efficient.

6.1.2 The kTuple Frequency Based method

The kTuple Frequency Based method is based on the hypothesis that a DNA sequence which has been horizontally transferred and incorporated into the genome occurs rarely. Due to the experiences with the abandoned implementations it was decided to decrease the length of the ktuples. The inquiry on which length to consider ended up on 10. The decision for using this length is questionable, since longer ktuples would be preferable. One reason for using this length was that the number of different ktuples to handle is approximately 1 million, which is a small enough number for keeping all of the ktuples in a dictionary. Another reason was the fact that with a length of 12 approximately 250 000 ktuples occurred only 1 time in the cow genome. It is probable that each of them will occur several times in the rice genome, which would have given an enormous amount of alignments. Performing alignments are time consuming as the following example shows.

In this thesis, the aligning of the expansions of the ktuples satisfying the LPHF criteria was performed. There were 4 ktuples meeting this criteria, and 3666 alignments were executed. With the current implementation, this execution took 69 minutes. It is known that the 1996 search-tuples occur < 20 times each in the cow genome, but for all that is known they might occur hundreds or perhaps thousands of times in the rice genome. A reduction to 248 ktuples was described in chapter 5.6.1. The average occurrences of these ktuples gives 248 000 alignments. By obtaining the same proportional increase of alignments and also the proportion of increase in number

of ktuples, this gives approximately 2 million alignments. With the current implementation this would take 26 days to execute.

The substitution matrix used as a scoring matrix in the alignments was based on the Kimura two-parameter model. The values chosen were inspired by those used by Rouchka (2007). By changing these values the alignments would most likely have been different, and those achieving the highest score would perhaps also have been different.

In accordance to the hypothesis stated the ktuples having a low frequency were found. The limit for a low frequency was decided to be < 20, based on interpretation of the plots in Figure 5.7. A verification of how appropriate this limit value is should be calculated on the basis of all frequencies. In this thesis further decrease of the frequencies were applied, which ended with those satisfying the LPHF criteria. All this reduction implies lower chances for any results to be achieved by the approach accomplished in this thesis.

The simulation that was performed to establish the discovery rate of the kTuple Frequency Based method was based on using all the 1996 search-tuples with frequencies < 20. As shown in Table 5.6, if a horizontally transferred sequence is 1000 base long, this method has a 31 % chance of discovering it. The simulation method does not take into account the mutations that naturally occur over time. The mutations rates should reflect the scoring matrix used in the alignments. If these mutations had been considered, more exact hit rates for a sequence of a certain length would be achieved.

The extension of the ktuples are done simply by retrieving the bases in positions 100 before and after the respective ktuple for all occurrences of it in both cow and rice. Other ways of expanding the ktuple could have been more appropriate. For instance by extending the ktuple in each direction, base by base, or perhaps a few bases at a time, and for each of these steps perform an alignment, better alignments would perhaps have been found. The way the expansion is done now, there is no guarantee that the ktuple is part of the alignment, or that it is aligned without a gap.

In the abandoned implementations, the reverse complementary of a sequence or ktuple was handled explicitly. In the kTuple Frequency Based method this has been left out. This is an unfortunate decision due to the influence the reverse complementary ktuples have on the total distribution of the ktuples. The frequencies found in this thesis only reflect the distribution on the forward strand of the genome. Including the frequencies on the reverse strand

might lead to different results.

6.2 Conclusion

For a horizontally transferred DNA sequence to be passed on to the next generation by the means of vertical inheritance, "the transfer must be made within the germ line of the new host" (Kurland, 2000). Nielsen (2006) performed experiments that found iv injected DNA in the gonad of Atlantic salmon, and Waters (2001) performed an experiment showing that the ovary cells of Chinese hamsters may experience horizontal transfers from *E. coli* by conjugation.

In this master thesis I have proposed a method called the kTuple Frequency Based method, which is a prototype method for detecting DNA from feed which has been horizontally transferred and incorporated into the genomes of eukaryotes. This method is based on the frequency of ktuples with length 10 in cow. It can be shortly summarized in these steps: The hypothesis that a DNA sequence which is horizontally transferred and incorporated into the genome occurs rarely is assumed. Ktuples with a low frequency are found, and several steps of reducing the number of search-tuples are carried out before these are expanded. All expansions of a ktuple from cow are aligned with all in rice. The last step is to use the sequences in the highest scoring alignments as queries in BLAST searches to see if significant matches are achieved.

In retrospect, more effort should have been put into achieving the best possible results rather than emphazising efficient execution time. By planning and preparing a search for HGT events thoroughly, an execution time of several days may be acceptable.

A bioinformatics method that guarantees the disclosure of a horizontal transfer event has yet to be created. A discovery rate as the one the kTuple Frequency Based method is simulated to have is promising. The results accomplished in this thesis is based on the 4 ktuples meeting the LPHF criteria. This, in addition to the fact that the genome sequences used as material are outdated, finding one cow sequence that actually achieves a significant match with rice in a BLAST search is quite an achievement. A downside is that the part of the sequence that achieved the match could not be concluded a horizontal transfer event.

The performance of the kTuple Frequency Based method in this thesis have been subject to several steps of reduction. I believe that if these steps were omitted, better results would have been achieved. I also believe that the proposed method can be generalized, and that with further improvements could turn out to be a worthy method for detecting HGT events.

6.3 Further work

Improvements are necessary in order to make the kTuple Frequency Based method a final approach. As presented in the previous section, the realization of the kTuple Frequency Based method in this thesis has been a prototype involving several steps of reduction. In a final approach based on this method some, or perhaps all, of these steps should be dropped. The different matters previously discussed, such as setting the frequency limit for the search-tuples at 20, should be further investigated. Perhaps it would be a good idea to set this limit based on calculations involving all the frequencies. Another improvement to the kTuple Frequency Based method is to increase the length of the ktuples. When the length of a ktuple is > 15, the original idea of using a close relative as a filter becomes possible. In this thesis the BLAST searches were performed manually. In a final version this should be implemented as an automatic task. All these improvements should be put together in one program.

The abandoned methods represent implementations inadequate to give results. As mentioned in the discussion, the basis algorithm if pursued in a different manner, could possibly achieve desired results. With several improvements these could form the basis for further work.

Bibliography

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997), 'Gapped BLAST and PSI-BLAST: a new generation of protein database search programs', *Nucleic Acids Research* **25**, 3389–3402.

ambion the rna company (2007), 'Dna and rna molecular weights and conversions', *http://www.ambion.com/techlib/append/na_mw_tables.html* .

Andersson, J. O. (2005), 'Lateral gene transfer in eukaryotes', *Cellular and Molecular Life Sciences* **62**, 1182–1197.

applied biosystems (2007), 'Essentials of real time pcr', *http://docs.appliedbiosystems.com/pebiodocs/00105622.pdf* .

Beiko, R. G., Harlow, T. J. and Ragan, M. A. (2005), 'Highways of gene sharing in prokaryotes', *Proceedings of the National Academy of Sciences* **102**, 14332–14337.

biopython Project (2006*a*), 'Biopython', *http://biopython.org/wiki/Documentation* .

biopython Project (2006*b*), 'Biopython api', *http://biopython.org/DIST/docs/api/* .

Blattner, F. R., III, G. P., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. and Shao, Y. (1997), 'The complete genome sequence of escherichia coli k-12', *Science* **277**, 1453–1462.

Brown, J. R. (2003), 'Ancient horizontal gene transfer', *Nature Reviews* **4**, 121–132.

Dayrat, B. (2003), 'The roots of phylogeny: How did Haeckel build his trees?', *Systematic Biology* **54**, 515–527.

de la Cruz, F. and Davies, J. (2000), 'Horizontal gene transfer and the origin of species: lessons from bacteria', *Trends in microbiology* **8**, 128–133.

Delwiche, C. F. and Palmer, J. D. (1996), 'Rampant horizontal transfer and duplication of rubisco genes in eubacteria and plastids', *Molecular Biology and Evolution* **13**, 873–882.

Doolittle, W. F. (1998), 'You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes', *Trends in Genetics* **14**, 307–311.

Doolittle, W. F. (1999), 'Phylogenetic classification and the universal tree', *Science* **284**, 2124–2128.

Doolittle, W. F., Nesbø, C. L., Bapteste, E. and Zhaxybayeva, O. (in press), Lateral gene transfer.

Dufraigne, C., Fertil, B., Lespinats, S., Giron, A. and Deschavanne, P. (2005), 'Detection and characterization of horizontal transfers in prokaryotes using genomic signature', *Nucleic Acid Research* **33**, ?–?

eGenix.com Software (2006), 'mxTextTools - Fast Text Manipulation Tools for Python', *http://www.egenix.com/files/python/mxTextTools.html* .

Einspanier, R., Klotz, A., Kraft, J., Aulrich, K., Poser, R., Schwägele, F., Jahreis, G. and Flachowsky, G. (2001), 'The fate of forage plant dna in farm animals: a collaborative case-study investigating cattle and chicken fed recombinant plant material', *European Food Research and Technology* **212**, 129–134.

Einspanier, R., Lutz, B., Rief, S., Berezina, O., Zverlov, V., Schwarz, W. and Mayer, J. (2004), 'Tracing residual recombinant feed molecules during digestion and rumen bacterial diversity in cattle fed transgene maize', *European Food Research and Technology* **218**, 269–273.

Eisen, J. A. (2000), 'Horizontal gene transfer among microbial genomes: insights from complete genome analysis', *Current Opinion in Genetics & Development* **10**, 606–611.

encyclopaedia Britannica (2007*a*), 'atomic mass', *http://www.britannica.com/eb/article-9010130* .

encyclopaedia Britannica (2007*b*), 'Charles darwin', *http://www.britannica.com/eb/article-9109642* .

encyclopaedia Britannica (2007*c*), 'polymerase chain reaction', *http://www.britannica.com/eb/article-9002524* .

Feng-Guang, G., Wen-Sheng, S., Ying-Lin, C., Li-Ning, Z., Jing, S., Hua-Fen, L. and Shi-Kun, Y. (1998), 'Hbx-dna probe preparation and its application in study of hepatocarcinogenesis', *World Journal of Gastroenterology* **4**, 320–322.

Garcia-Vallé, S., Romeu, A. and Palau, J. (2000), 'Horizontal gene transfer in bacterial and archaeal complete genomes', *Genome Research* **10**, 1719–1725.

glimmer mobile gene-finding system (2007), 'Glimmer', *http://www.cbcb.umd.edu/software/glimmer/* .

Gogarten, J. P., Murphy, R. D. and Oldenzenski, L. (1999), 'Horizontal gene transfer: Pitfalls and promises', *The Biological Bulletin* **196**, 359–362.

Gogarten, J. P. and Townsend, J. P. (2005), 'Horizontal gene transfer, genome innovation and evolution', *Nature Reviews* **3**, 679–687.

Jonas, D. A., Elmadfa, I., Engel, K.-H., Heller, K., König, A., Müller, D., Narbonne, J., Wackernagel, W. and Kleiner, J. (2001), 'Safety considerations of dna in food', *Annuals of Nutrition and Metabolism* **45**, 235–254.

Khalsa, G. (2007), 'Southern blotting'. *http://lifesciences.asu.edu/resources/mamajis/southern/southern.html*.

Kimura, M. (1980), 'A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences', *Journal of Molecular Evolution* **16**, 111–120.

Koonin, E. V., Makarova, K. S. and Aravind, L. (2001), 'Horizontal gene transfer in prokaryotes: quantification and classification', *Annual Reviews* **55**, 709–742.

Kurland, C. G. (2000), 'Something for everyone. horizontal gene transfer in evolution', *European Molecular Biology Organization* **1**, 92–95.

Kurland, C. G., Canback, B. and Berg, O. G. (2003), 'Horizontal gene transfer: A critical view', *Proceedings of the National Academy of Sciences* **100**, 9658–9662.

Lawrence, J. G. and Ochman, H. (1997), 'Amelioratin of bacterial genomes: Rates of change and exchange', *Journal of Molecular Evolution* **44**, 383–397.

Lawrence, J. G. and Ochman, H. (1998), 'Molecular archaeology of the escherichia coli genome', *Proceedings of the National Academy of Sciences* **95**, 9413–9417.

Licht, T. R. and Wilcks, A. (2006), 'Conjugative gene transfer in the gastrointestinal environment', *Advances in applied microbiology* **58**, 77–95.

L.Theobald, D. (2007), '29+ evidences for macroevolution: The scientific case for common descent', *http://www.talkorigins.org/faqs/comdesc/CDhierarchy.html#CDfig* .

McDonald, D. (2007), 'Microsatellite dna: the half-page intro', *http://www.uwyo.edu/dbmcd/lab/msatintro.html* .

Moore, J. (2007), 'The Boyer-Moore Fast String Searching Algorithm', *http://www.cs.utexas.edu/users/moore/best-ideas/string-searching/* .

Nakamura, Y., Itoh, T., Matsuda, H. and Gojobori, T. (2004), 'Biased biological functions of horizontal transferred genes in prokaryotic genomes', *Nature Genetics* **36**, 760–766.

NCBI (2006*a*), 'National Center for Biotechnology Information', *http://www.ncbi.nlm.nih.gov* .

NCBI (2006*b*), 'Ncbi ftp site', *ftp://ftp.ncbi.nih.gov* .

NCBI (2007), 'Basic Local Alignment Search Tool', *http://www.ncbi.nlm.nih.gov/BLAST* .

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A., McDonald, L., Utterback, T. R., Malek, J. A., Linher, K. D., Garrett, M. M., Stewart, A. M., Cotton, M. D., Pratt, M. S., Phillips, C. A., Richardson, D., Heidelberg, J., Sutton, G. G., Fleischmann, R. D., Eisen, J. A., White, O., Salzberg, S. L., Smith, H. O., Venter, J. C. and Fraser, C. M. (1999), 'Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *thermotoga maritima*', *Nature* **399**, 323–329.

Neu, H. C. (1992), 'The crisis in antibiotic resistance', *Science* **257**, 1064–1073.

Nielsen, C. R. (2006), Tracing and characterisation of dietary DNA - applications in Atlantic salmon, PhD thesis, Faculty of Mathematics and Natural Science, University of Oslo.

Nielsen, C. R., Berdal, K. G., Bakke-McKellep, A. M. and Holst-Jensen, A. (2005), 'Dietary dna in blood and organs of atlantic salmon (*salmo salar* l.)', *European Food Research and Technology* **221**, 1–8.

Nielsen, C. R., Holst-Jensen, A., Løvseth, A. and Berdal, K. G. (2006), 'Persistence and distribution of intravenously injected dna in blood and organs of atlantic salmon (*salmo salar* l.)', *European Food Research and Technology* **222**, 258–265.

Ning, Z., Cox, A. J. and Mullikin, J. C. (2001), 'SSAHA: A fast search method for large DNA databases', *Genome Research* **11**, 1725–1729.

pasteur institute (2007), 'Fluorescent in situ hybridization (FISH)'. *http://www.pasteur.fr/recherche/unites/biophyadn/e-fish.html*.

Phipps, R. H., Deaville, E. R. and Maddison, B. C. (2003), 'Detection of transgenic and endogenous plant dna in rumen fluid, duodenal digesta, milk, blood, and feces of lactating dairy cows', *Journal of Dairy Science* **86**, 4070–4078.

Python (2006), 'The python programming language'. *http://www.python.org/*.

Ragan, M. A. (2001), 'On surrogate methods for detecting lateral gene transfer', *FEMS Microbiology Letters* **201**, 187–191.

R.Bordenstein, S. and Reznikoff, W. S. (2005), 'Mobile dna in obligate intracellular bacteria', *Nature Reviews* **3**, 688–699.

Richardson, A. O. and Palmer, J. D. (2007), 'Horizontal gene transfer in plants', *Journal of Experimental Botany* **58**, 1–9.

Rouchka, E. C. (2007), 'Aligning dna sequences using dynamic programming'. *http://www.acm.org/crossroads/xrds13-1/dna.html*.

Salzberg, S. L., White, O., Peterson, J. and Eisen, J. A. (2001), 'Microbial genes in the human genome: Lateral transfer or gene loss?', *Science* **292**, 1903–1906.

Schubbert, R., Hohlweg, U., Renz, D. and Doerfler, W. (1998), 'On the fate of orally ingested foreign dna in mice: chromosomal association and placental transmission to the fetus', *Molecular Genetics and Genomics* **259**, 569–576.

Schubbert, R., Lettmann, C. and Doerfler, W. (1994), 'Ingested foreign (phage m13) dna survives transiently in the gastrointestinal tract and enters the bloodstream of mice', *Molecular Genetics and Genomics* **242**, 495–504.

Schubbert, R., Renz, D., Schmitz, B. and Doerfler, W. (1997), 'Foreign (m13) dna ingested by mice reaches peripheral leukocytes, spleen, and liver via the intestinal wall mucosa and can be covalently linked to mouse dna', *Proceedings of the National Academy of Sciences of the United States of America* **94**, 961–966.

Shi, S.-Y., Cai, X.-H. and fu Ding, D. (2005), 'Identification and categorization of horizontally transferred genes in prokaryotic genomes', *Acta Biochimica et Biophysica Sinica* **37**, 561–566.

Smets, B. F. and Barkay, T. (2005), 'Horizontal gene transfer: perspectives at a crossroads of scientific disciplines', *Nature Reviews* **3**, 675–678.

Snel, B., Huynen, M. A. and Dutilh, B. E. (2005), 'Genome evolution and the nature of genome evolution', *Annual Reviews Microbiology* **59**, 191–209.

Sutton, W. S. (1903), 'The chromosomes in heredity', *The Biological Bulletin* **4**, 231–250.

TIGR (2007), 'Thermotoga maritima', *http://www.tigr.org/tdb/CMR/btm/htmls/Background.html* .

Waters, V. L. (2001), 'Conjugation between bacterial and mammalian cells', *Nature Genetics* **29**, 375–376.

Woese, C. (1998), 'The universal ancestor', *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6854–6859.

Xiong, J. (2006), *Essential Bioinformatics*, Cambridge University Press.

Appendix A

Index of the source files

A.1 Python-source

aligning_expanded2.py	Performs the alignments for the expansions of the ktuples meeting the LP or LPHF criteria. The output is saved in kjoring_aligning_exp2_met1.txt and kjoring_aligning_exp2_met2.txt by the >> command in linux.
downsizing_expanded.py	A means for collecting information on which the choices for the final reduction of the searchtuples were made. Have been modified several times. Results from the different runs are saved in kjoring_ds_ex.txt, kuttet_ned.txt, mindre_enn_6_expanded.txt and m_e_6_exp_sort.txt by the >> command in linux.
expand_ktuples.py	Extracts the expansions of the chosen ktuples. These are saved in on file for cow, utvidelser.ku and one for rice, utvidelser.ris.
finne_ant_baser.py	Counts the number of bases in the sequences.

interesting_tuples.py	Treats the .res files. Finds the amount of ktuples having one frequency and also those with frequencies < 20. The result is saved in antall_10_TuplerKu.sok.
ku_tupleOversikt.py	Finds all ktuples and their fequencies in the cowfiles. Saves the result in antall_10_TuplerKu.res.
kuTupleDistribusjonSortering.py	Sorts the .diagram files. The result is saved in antall_X_TuplerKu_sortert.diagram. X=[8,10,12].
rice_search.py	The first method for searching in rice
rice_search2.py	The second method for searching in rice
rice_search3.py	The third method for searching in rice. The result is saved in the antall_10_TuplerKu.found.
search_results.py	Extracts info from .found files
simulering.py	Performs the simulation of the kTuple Frequency Based method. The result is manually saved in simuleringsResultat.txt
test_finnes_sek_i_ku.py	Finds where in the genome sequences a subsequence is located. The result is saved in the kjoring_test_finnes_sek_i_ku.txt by the >> command in linux.

A.2 R-source

lage_grafer.r	Creates the "Frequency vs. number with frequency"-plots

A.3 Result-files

antall_X_TuplerKu.data	For the cow-files: <*frequency*> <*number with this frequency*> <*ktuples*> The ktuples format: ['ktuple1','ktuple2' ...]. X=[8,10,12].
antall_X_TuplerKu.diagram	For the cow-files: <*frequency*> <*number with this frequency*>. X=[8,10,12].
antall_10_TuplerKu.found	For the rice-files: <*ktuple*> <*minimum frequency*>.
antall_X_TuplerKu.res	For the cow-files: <*ktuple*> <*frequency*>. X=[8,10,12].
antall_10_TuplerKu.sok	The search-tuples from cow: <*frequency*> <*number with this frequency*> <*ktuples*> The ktuples format: *ktuple1 ktuple2 ktuple3 ...*
antall_X_TuplerKu_sortert.diagram	For the cow files: <*frequency*> <*number with this frequency*> sorted on the frequency. Used to create the distribution-plots. X=[8,10,12].
kjoring_aligning_exp2_met1.txt	Contains the top 10 alignments for the LP criteria.
kjoring_aligning_exp2_met2.txt	Contains the top 10 alignments for the LPHF criteria.
kjoring_ds_ex.txt	The search-tuples $\leq$ 10: <*ktuple*> <*frequency cow*> <*frequency rice*>

kjoring_test_finnes_sek_i_ku.txt	Contains the result from the test run to verify that the nr.1, nr.2 and nr.3 sequences are in the cow-files. Shows which files, sequences and positions they are in.
kuttet_ned.txt	The search-tuples ≤ 8: <*ktuple*> <*frequency cow*> <*frequency rice*>
m_e_6_exp_sort.txt	The search-tuples ≤ 5: <*ktuple*> <*frequency cow*> <*frequency rice*> sorted on the frequency on cow.
metode_1_tupler.align	The ktuples meeting the LP criteria
metode_2_tupler.align	The ktuples meeting the LPHF criteria
mindre_enn_6_expanded.txt	The search-tuples ≤ 5: <*ktuple*> <*frequency cow*> <*frequency rice*>
simuleringsResultat.txt	The result from 10 different runs of the simulation.
utvidelser.ku	For cow: <*search-tuple*> <*expansions*> The expansions format: *exp1 exp2 exp3 ...*
utvidelser.ris	For rice: The expansions format: *exp1 exp2 exp3 ...*